

Terry D. Oswalt
Editor-in-Chief

Gerard Gilmore
Volume Editor

Planets, Stars and Stellar Systems

VOLUME 5

Galactic Structure and Stellar Populations

Planets, Stars and Stellar Systems

Galactic Structure and Stellar Populations

Terry D. Oswalt (Editor-in-Chief)

Gerard Gilmore (Volume Editor)

Planets, Stars and Stellar Systems

Volume 5: Galactic Structure and Stellar Populations

With 452 Figures and 38 Tables



Springer Reference

Editor-in-Chief

Terry D. Oswalt
Department of Physics & Space Sciences
Florida Institute of Technology
University Boulevard
Melbourne, FL, USA

Volume Editor

Gerard Gilmore
Institute of Astronomy
Cambridge University, UK

ISBN 978-94-007-5611-3 ISBN 978-94-007-5612-0 (eBook)
ISBN 978-94-007-5613-7 (print and electronic bundle)
DOI 10.1007/978-94-007-5612-0

This title is part of a set with
Set ISBN 978-90-481-8817-8
Set ISBN 978-90-481-8818-5 (eBook)
Set ISBN 978-90-481-8852-9 (print and electronic bundle)

Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012953926

© Springer Science+Business Media Dordrecht 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Series Preface

It is my great pleasure to introduce “Planets, Stars, and Stellar Systems” (PSSS). As a “Springer Reference”, PSSS is intended for graduate students to professionals in astronomy, astrophysics and planetary science, but it will also be useful to scientists in other fields whose research interests overlap with astronomy. Our aim is to capture the spirit of 21st century astronomy – an empirical physical science whose almost explosive progress is enabled by new instrumentation, observational discoveries, guided by theory and simulation.

Each volume, edited by internationally recognized expert(s), introduces the reader to a well-defined area within astronomy and can be used as a text or recommended reading for an advanced undergraduate or postgraduate course. Volume 1, edited by Ian McLean, is an essential primer on the tools of an astronomer, i.e., the telescopes, instrumentation and detectors used to query the entire electromagnetic spectrum. Volume 2, edited by Howard Bond, is a compendium of the techniques and analysis methods that enable the interpretation of data collected with these tools. Volume 3, co-edited by Linda French and Paul Kalas, provides a crash course in the rapidly converging fields of stellar, solar system and extrasolar planetary science. Volume 4, edited by Martin Barstow, is one of the most complete references on stellar structure and evolution available today. Volume 5, edited by Gerard Gilmore, bridges the gap between our understanding of stellar systems and populations seen in great detail within the Galaxy and those seen in distant galaxies. Volume 6, edited by Bill Keel, nicely captures our current understanding of the origin and evolution of local galaxies to the large scale structure of the universe.

The chapters have been written by practicing professionals within the appropriate sub-disciplines. Available in both traditional paper and electronic form, they include extensive bibliographic and hyperlink references to the current literature that will help readers to acquire a solid historical and technical foundation in that area. Each can also serve as a valuable reference for a course or refresher for practicing professional astronomers. Those familiar with the “Stars and Stellar Systems” series from several decades ago will recognize some of the inspiration for the approach we have taken.

Very many people have contributed to this project. I would like to thank Harry Blom and Sonja Guerts (Sonja Japenga at the time) of Springer, who originally encouraged me to pursue this project several years ago. Special thanks to our outstanding Springer editors Ramon Khanna (Astronomy) and Lydia Mueller (Major Reference Works) and their hard-working editorial team Jennifer Carlson, Elizabeth Ferrell, Jutta Jaeger-Hamers, Julia Koerting, and Tamara Schineller. Their continuous enthusiasm, friendly prodding and unwavering support made this series possible. Needless to say (but I’m saying it anyway), it was not an easy task shepherding a project this big through to completion!

Most of all, it has been a privilege to work with each of the volume Editors listed above and over 100 contributing authors on this project. I’ve learned a lot of astronomy from them, and I hope you will, too!



January 2013

Terry D. Oswalt
General Editor

Preface to Volume 5

Studies of Stellar Populations have developed over the half-century since the term became popular, at the Vatican Conference of 1957, to become the critical methodology for precision studies of the evolution of star formation with time and location, galaxy assembly, the evolution of the chemical elements and their return to the inter-Galactic medium, and the spatial distribution of dark matter. None of those applications was strongly represented in the Vatican Conference as we discuss them today – rather Stellar Populations and Galactic Structure is a modern, vibrant, and fast-changing subject, building on those big questions which motivate substantial aspects of astronomy today.

The chapters in this volume show both the history of the discoveries and the evolution in complexity of the subject, from initial concern very largely with observationally based determinations of the three-dimensional spatial and kinematic distributions of the various stellar populations, defined by stellar type, and the inter-stellar medium. Following that groundwork, it is perhaps our realization that the inter-stellar and the intergalactic medium hold most of the Universe's baryons, and much complex chemistry, that have contributed immensely toward recent understanding – the context of galaxies as “Island Universes” remains a surprisingly good descriptor of observations; yet we know all structures in the Universe live in a broader cosmological context.

Naturally, newly enabled technologies, especially in high-energy studies, and in high spatial resolution imaging/kinematic studies of both stars and gas, have no early counterpart. The several chapters in this volume describing the extension of our knowledge beyond the optical and low-resolution HI 21-cm radio studies highlight impressive progress, illustrating a paradigm shift in our Galactic view.

The high intrinsic interest in understanding the implications of kinematics for mass determinations and for Galaxy evolution has been an active field throughout the twentieth century, as is still the case, and merits extensive discussion. The chapter on “History of Dark Matter in Galaxies” presents the discovery of our current concept of Dark Matter reminds us that some radical discoveries mature only slowly into our consciousness, others leap forth fully formed.

Stellar chemical abundance determination and interpretation is now a major science, with quantitative high-resolution spectroscopy on large telescopes driving impressive progress. In consequence, this features very largely in the present volume, representing its critical role in modern astrophysics. With a range of 5dex in measured abundances being available for analysis, we have now available a probe of the Universe from the very earliest star formation to the present day.

The stellar Initial Mass Function was as fundamental and as much debated in the earliest days of the study of the Galaxy as now, with the chapter by van Rhijn, written in 1959, published in the 1965 volume, providing a direct link to the pioneering days of the subject. Comparison of van Rhijn's Table 1 with chapter 4 illustrates the impressive advances in astronomical technology, data, and modeling sophistication that a half-century has bought, while reminding us that some fundamental questions remain open for debate.

Several chapters in the present volume develop the theme of determination of the three-dimensional structure of the Milky Way Galaxy, a challenge in which substantial progress

has been made – especially through digital star count data in both optical and infrared wavelengths – and where we anticipate a revolution over the next decade with the Gaia mission.

This is a natural time to take stock of what has been achieved with ground-based optical and near-infrared imaging surveys, the beginnings of large spectroscopic surveys, the first generations of space-based thermal-infrared surveys, superb multiwavelength complementary data, including both photon and particle messengers, and when a fundamental calibration of the distance scale – that holy grail of astrophysics – is finally within sight in the coming decade.

Gerard Gilmore
UK

Editor-in-Chief



Dr. Terry D. Oswalt

Department Physics & Space Sciences
Florida Institute of Technology
150 W. University Boulevard
Melbourne, Florida 32901
USA
E-mail: toswalt@fit.edu

Dr. Oswalt has been a member of the Florida Tech faculty since 1982 and was the first professional astronomer in the Department of Physics and Space Sciences. He serves on a number of professional society and advisory committees each year. From 1998 to 2000, Dr. Oswalt served as Program Director for Stellar Astronomy and Astrophysics at the National Science Foundation. After returning to Florida Tech in 2000, he served as Associate Dean for Research for the College of Science (2000–2005) and interim Vice Provost for Research (2005–2006). He is now Head of the Department of Physics & Space Sciences. Dr. Oswalt has written over 200 scientific articles and has edited three astronomy books, in addition to serving as Editor-in-Chief for the six-volume Planets, Stars, and Stellar Systems series.

Dr. Oswalt is the founding chairman of the Southeast Association for Research in Astronomy (SARA), a consortium of ten southeastern universities that operates automated 1-meter class telescopes at Kitt Peak National Observatory in Arizona and Cerro Tololo Interamerican Observatory in Chile (see the website www.saraobservatory.org for details). These facilities, which are remotely accessible on the Internet, are used for a variety of research projects by faculty and students. They also support the SARA Research Experiences for Undergraduates (REU) program, which brings students from all over the U.S. each summer to participate one-on-one with SARA faculty mentors in astronomical research projects. In addition, Dr. Oswalt secured funding for the 0.8-meter Ortega telescope on the Florida Tech campus. It is the largest research telescope in the State of Florida.

Dr. Oswalt's primary research focuses on spectroscopic and photometric investigations of very wide binaries that contain known or suspected white dwarf stars. These pairs of stars, whose separations are so large that orbital motion is undetectable, provide a unique opportunity to explore the low luminosity ends of both the white dwarf cooling track and the main sequence; to test competing models of white dwarf spectral evolution; to determine the space motions, masses, and luminosities for the largest single sample of white dwarfs known; and to set a lower limit to the age and dark matter content of the Galactic disk.

Volume Editor



Gerard Gilmore
Institute of Astronomy
Cambridge University
UK

Gerard Gilmore is professor of experimental philosophy at the Institute of Astronomy, Cambridge University, UK. He completed his Ph.D. in New Zealand, then worked for 5 years at the Royal Observatory Edinburgh on the first digitization of the Sky Survey photographic plates being obtained at the newly commissioned UK Schmidt Telescope, before moving to Cambridge.

He is a member of a very large number of local, national, and international review, policy, and assessment committees, has been UK Scientific Representative on the Council of the European Southern Observatory; is Scientific Coordinator of OPTICON, the EC Optical-Infrared Coordination Network for Astronomy; is UK Principal Investigator for the ESA Gaia mission; and is Co-PI of the Gaia-ESO Public Spectroscopic Survey.

The early photographic star count studies led to significantly improved determination of the low-luminosity stellar mass function, with implications for the possible significance of baryonic dark matter. These star-count studies independently discovered the Galactic Thick Disk.

Adding kinematics and spectroscopy to these star counts made possible an improved determination of the density and distribution of Dark Matter near the Sun, with a result which remains the accepted value 30 years later.

Later studies involved determination of the chemical abundance distributions and kinematics in the various stellar populations and their interpretation. Discovery of the Sagittarius dwarf galaxy, the “smoking gun” evidence for ongoing Galactic assembly, became the beginning of many studies of Galactic satellite galaxies, their implications for early galaxy formation, and the spatial distribution of Cold Dark Matter. These studies continue, with some 600 publications to date.

Table of Contents

Series Preface	v
Preface to Volume 5	vii
Editor-in-Chief	ix
Volume Editor	xi
List of Contributors	xv

Volume 5

1 Stellar Populations	1
<i>Rosemary F. G. Wyse</i>	
2 Chemical Abundances as Population Tracers	21
<i>Poul Erik Nissen</i>	
3 Metal-Poor Stars and the Chemical Enrichment of the Universe	55
<i>Anna Frebel · John E. Norris</i>	
4 The Stellar and Sub-Stellar Initial Mass Function of Simple and Composite Populations	115
<i>Pavel Kroupa · Carsten Weidner · Jan Pflamm-Altenburg · Ingo Thies · Jörg Dabringhausen · Michael Marks · Thomas Maschberger</i>	
5 The Galactic Nucleus	243
<i>Fulvio Melia</i>	
6 The Galactic Bulge	271
<i>R. Michael Rich</i>	
7 Open Clusters and Their Role in the Galaxy	347
<i>Eileen D. Friel</i>	
8 Star Counts and the Nature of the Galactic Thick Disk	393
<i>Yuzuru Yoshii</i>	
9 The Infrared Galaxy	447
<i>Ed Churchwell · Robert A. Benjamin</i>	

10	Interstellar PAHs and Dust	499
	<i>A. G. G. M Tielens</i>	
11	Galactic Neutral Hydrogen	549
	<i>John M. Dickey</i>	
12	High-Velocity Clouds	587
	<i>Bart P. Wakker · Hugo van Woerden</i>	
13	Magnetic Fields in Galaxies	641
	<i>Rainer Beck · Richard Wielebinski</i>	
14	Astrophysics of Galactic Charged Cosmic Rays	725
	<i>Antonella Castellina · Fiorenza Donato</i>	
15	Gamma-Ray Emission of Supernova Remnants and the Origin of Galactic Cosmic Rays	789
	<i>F. A. Aharonian</i>	
16	Galactic Distance Scales	829
	<i>Michael W. Feast</i>	
17	Globular Cluster Dynamical Evolution	879
	<i>Melvyn B. Davies</i>	
18	Dynamics of Disks and Warps	923
	<i>J. A. Sellwood</i>	
19	Mass Distribution and Rotation Curve in the Galaxy	985
	<i>Yoshiaki Sofue</i>	
20	Dark Matter in the Galactic Dwarf Spheroidal Satellites	1039
	<i>Matthew Walker</i>	
21	History of Dark Matter in Galaxies	1091
	<i>Virginia Trimble</i>	
	Index	1119

List of Contributors

F. A. Aharonian

Dublin Institute for Advanced Studies
Dublin
Ireland
and
Max-Planck-Institut für Kernphysik
Heidelberg
Germany

Rainer Beck

Max-Planck-Institut für Radioastronomie
Bonn
Germany

Robert A. Benjamin

Department of Physics
University of Wisconsin - Whitewater
Whitewater, WI
USA

Antonella Castellina

Osservatorio Astrofisico di Torino
Istituto Nazionale di Astrofisica
Torino
Italy

Ed Churchwell

Department of Astronomy
University of Wisconsin
Madison, WI
USA

Jörg Dabringhausen

Argelander-Institut für Astronomie
Universität Bonn
Bonn
Germany

Melvyn B. Davies

Lund Observatory
Lund
Sweden

John M. Dickey

School of Mathematics and Physics
University of Tasmania
Hobart, TAS
Australia

Fiorenza Donato

Dipartimento di Fisica
Universita' di Torino
Torino
Italy

Michael W. Feast

Astronomy Department and Astrophysics
Cosmology and Gravity Centre
University of Cape Town and South African
Astronomical Observatory
Rondebosch
South Africa

Anna Frebel

Department of Physics
Massachusetts Institute of Technology &
Kavli Institute for Astrophysics and Space
Research
Cambridge, MA
USA

Eileen D. Friel

Department of Astronomy
Indiana University
Bloomington, Indiana
USA

Pavel Kroupa

Argelander-Institut für Astronomie
Universität Bonn
Bonn
Germany

Michael Marks

Argelander-Institut für Astronomie
Universität Bonn
Bonn
Germany

Thomas Maschberger

Institute of Astronomy
Cambridge
UK
and
Institut de Planétologie et d' Astrophysique
de Grenoble
Grenoble Cédex 9
France

Fulvio Melia

Department of Physics
Steward Observatory
and the Applied Math Program
The University of Arizona
Tucson, AZ
USA

Poul Erik Nissen

Department of Physics and Astronomy
University of Aarhus
Aarhus C
Denmark

John E. Norris

Research School of Astronomy and
Astrophysics
Australian National University
Canberra, ACT
Australia

Jan Pflamm-Altenburg

Argelander-Institut für Astronomie
Universität Bonn
Bonn
Germany

R. Michael Rich

Department of Physics and Astronomy
University of California
Los Angeles, CA
USA

J. A. Sellwood

Department of Physics and Astronomy
Rutgers
The State University of New Jersey
Piscataway, NJ
USA

Yoshiaki Sofue

Institute of Astronomy
The University of Tokyo
Tokyo
Japan
and
Department of Physics
Meisei University
Tokyo
Japan

Ingo Thies

Argelander-Institut für Astronomie
Universität Bonn
Bonn
Germany

A. G. G. M. Tielens

Leiden Observatory
Leiden University
Leiden
The Netherlands

Virginia Trimble

Department of Physics and Astronomy
University of California
Irvine, CA
USA

Bart P. Wakker

Supported by NASA and NSF; affiliated
with Department of Astronomy
University of Wisconsin
Madison, WI
USA

Matthew Walker

Harvard-Smithsonian Center for
Astrophysics
Cambridge, MA
USA
and
Hubble Fellow

Carsten Weidner

Scottish Universities Physics Alliance
(SUPA)
School of Physics and Astronomy
University of St. Andrews
North Haugh
St. Andrews
UK
and
Instituto de Astrofísica de Canarias
La Laguna (Tenerife)
Spain

Richard Wielebinski

Max-Planck-Institut für Radioastronomie
Bonn
Germany

Hugo van Woerden

Kapteyn Astronomical Institute
Rijksuniversiteit Groningen
Groningen
The Netherlands

Rosemary F. G. Wyse

Department of Physics and Astronomy
The Johns Hopkins University
Baltimore, MD
USA

Yuzuru Yoshii

Institute of Astronomy
School of Science
University of Tokyo
Tokyo
Japan

1 Stellar Populations

Rosemary F. G. Wyse


Department of Physics and Astronomy, The Johns Hopkins
University, Baltimore, MD, USA

1	<i>Introduction: Definitions of Populations I and II</i>	2
2	<i>Early Understanding of Populations I and II</i>	3
3	<i>Some Complexities: What Is Population II?</i>	7
3.1	Thick Disks	7
3.2	Bulges	9
3.2.1	Milky Way Galaxy	10
3.2.2	M31	12
3.3	Stellar Halo	13
3.3.1	Field Stars	13
3.4	Satellite Galaxies	14
3.4.1	Globular Clusters	15
4	<i>Cosmological Implications of the Properties of (Galactic) Stellar Populations</i> ...	15
	<i>Acknowledgments</i>	17
	<i>References</i>	18

Abstract: Stellar populations encode the star-formation history and chemical evolution of a system. This chapter reviews the development of the concept of stellar populations and the current understanding of the resolved stellar populations in the Local Group, with an emphasis on the older stars and serves to introduce the topics of several later chapters.

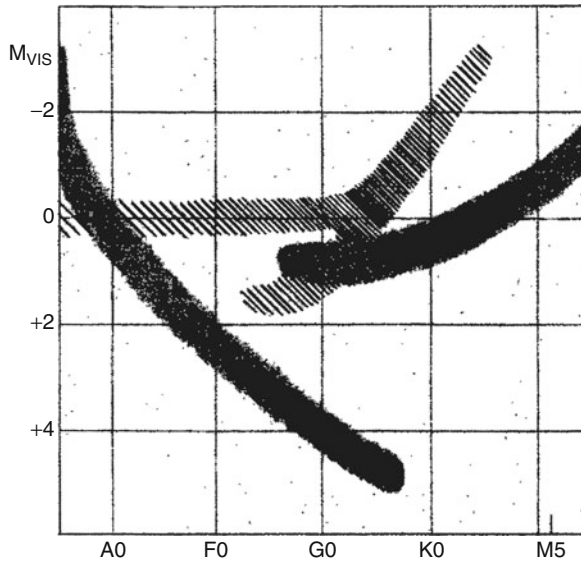
Keywords: Galaxies: formation, Galaxy: evolution, Galaxy: structure, Stars: abundances, Stars: kinematics

1 Introduction: Definitions of Populations I and II

The concept of stellar populations was introduced by Baade (1944), upon his resolution of the brightest stars in the central regions of M31 and in its satellite galaxies M32 and NGC 205. The stars were resolved in red-sensitive photographic plates, but not in blue-sensitive plates, confirming that they were *not* hot, luminous stars similar to those brightest stars found in the solar neighborhood. Baade identified these newly resolved (extraGalactic) stars as being the counterparts of the brightest stars in Galactic globular clusters. The differences between the stellar loci in the Hertzsprung-Russell diagram of solar-neighborhood stars and globular-cluster stars were illustrated by Baade in a figure reproduced here as  Fig. 1-1. This formed the basis for the separation of the known stars into two populations, with Population I being exemplified by the stars in the solar neighborhood, having luminous early-type stars and luminous M-type stars, while Population II – exemplified by stars in globular clusters – lacks luminous O- and B-type stars and also lacks luminous M-type stars. We now understand these differences as being due to the fact that stars in globular clusters are all older than the luminous blue stars in the solar neighborhood, which leads to a low-mass, red main-sequence turnoff, combined with the low metallicity of globular-cluster stars, which leads to a blue red-giant branch, hotter than the M-type stars.

A “stellar population” was then taken to mean a collection of stars with a well-defined locus on the H-R diagram, implying well-defined age and metallicity distributions. Indeed, we now understand the term Simple Stellar Population to mean a single generation of stars, with some initial mass function, all of the same chemical abundances and age. Stellar Population Synthesis is the process by which one combines Simple Stellar Populations in such a way as to match the observed photometry and spectroscopy of composite systems such as galaxies (including complications caused by the presence of dust).

The fact that stars of solar mass live for approximately the present age of the universe (and longer for lower-mass stars) offer the opportunity to use them to infer conditions at very early epochs. The surface chemical abundances are conserved, to first order, over a star’s lifetime, and the elemental abundance distributions of stars may be used to constrain the mass function of those stars that pre-enriched them, together with gas flows and the (in)homogeneity of star formation. Stars are (in general) collisionless and thus in a fixed potential conserve orbital quantities such as angular momentum; kinematic signatures of a “stellar population” persist even through the merging predicted in Λ CDM. The study of the “fossil record” from old stars nearby provides a complementary probe of the early stages of galaxy formation to that afforded by the study of high redshift objects. Analysis of resolved stars allows the breaking of degeneracies, such as that between age and metallicity, often encountered in the interpretation of the integrated properties of distant galaxies. Stars of different ages provide a series of snapshots of



■ Fig. 1-1


The schematic Hertzsprung-Russell diagram (spectral type, as a proxy for surface temperature, against absolute magnitude) used by Baade (1944) to illustrate the definitions of his two populations: The locus of Population I, stars in the solar neighborhood, is represented by the darker areas, while the locus of Population II, stars in globular clusters, is represented by the lighter hatched areas. Population II lacks the bright blue main-sequence stars of Population I and has a significantly bluer red-giant branch. These differences are due to, respectively, the older age of Population II stars and their lower metallicity (Fig. 1 of Baade (1944), used with permission)

one galaxy over time, compared to snapshots of different galaxies at different redshifts. Each approach provides unique insight, most powerful when combined. The present review focusses on resolved stellar populations, with an emphasis on older stars.

2 Early Understanding of Populations I and II

As Baade (1944) discussed in his paper, Oort (1926) had earlier shown that luminous, early-type (O and B) stars were essentially missing from his sample of “high-velocity” stars (referring to motion relative to the Sun), thus introducing kinematics into the concept of stellar populations: Population II stars are high-velocity stars. Baade also noted that local “subdwarfs” were high-velocity stars. These subdwarf stars are so-called since they lie below the normal main dwarf sequence in the absolute magnitude – color plane. Several lines of investigation in the 1950s (Roman 1955; Schwarzschild et al. 1955; Sandage and Eggen 1959) provided the explanation: the subdwarfs are bluer than “normal” stars of the same mass, rather than being subluminous, reflecting the fact that their atmospheres lack much of the opacity (at blue wavelengths) from elements heavier than helium, i.e., high-velocity subdwarfs are metal poor. How much bluer subdwarfs are can be quantified by the excess of U-band light relative to “normal” metallicity

stars, called the “UV excess,”¹ calibrated using spectroscopic determinations of metallicity (e.g., Wallerstein 1962; Carney 1979). Lower-metallicity stars have higher value of the UV excess. Such photometry-based estimates of metallicity can be obtained for fainter stars, in larger numbers compared to spectroscopic estimates obtained using the same exposure time, and thus are invaluable for surveys (provided the limitations of the calibration are treated properly).

The culmination of this early work, as applied to the Milky Way Galaxy, was the seminal paper by Eggen et al. (1962; ELS). These authors famously analyzed a sample of 221 nearby F/G turnoff (dwarf) stars with full 3D space motions and determined the relationship between UV excess and kinematics, using orbital eccentricity, vertical velocity, and angular momentum as the relevant parameters. There were three main reasons for their choice of F/G stars: they are long-lived so should trace all stages of Galaxy evolution, they are represented in both Population I and Population II, and the UV excess is calibrated for such stars. ELS argued that their data were consistent with a smooth correlation between metallicity and kinematics, in the sense that lower metallicity goes with higher-amplitude non-circular motions. Further, they proposed that metallicity should be a reliable proxy for age, leading to consistency with a smooth correlation between age and kinematics. Their plot showing the absolute value of the vertical component of space motion ($|W|$) against the UV excess ($\delta(U - B)$) is reproduced here in  Fig. 1-2; the right-hand ordinate axis shows the derived maximum height of each star, based on its present position and velocity and an adopted model potential for the Milky Way.

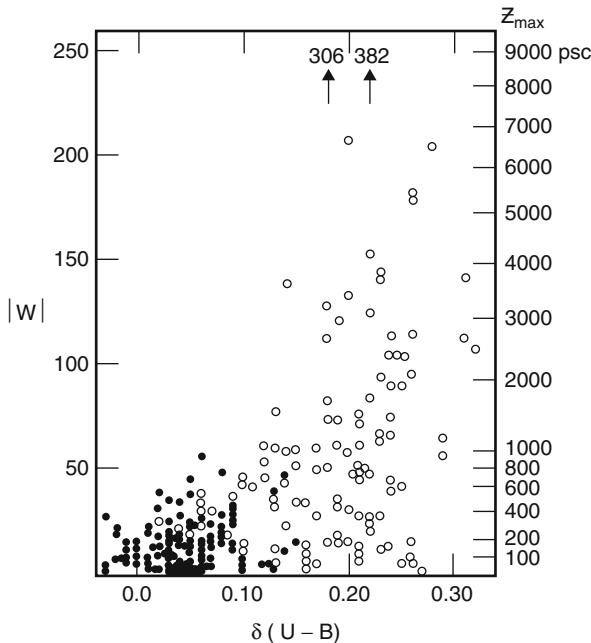


 Fig. 1-2

UV excess, which is higher for lower-metallicity stars, against the absolute value of velocity in the vertical direction (*left y-axis*) and derived maximum vertical height reached (*right y-axis*) for a sample of Galactic F/G turnoff stars (Fig. 5 from Eggen et al. (1962), used with permission)

¹Most usefully, the $(U - B)$ color excess, normalized to a star with $(B - V) = 0.6$ on the main sequence of the Hyades cluster (Sandage and Eggen 1959)

A higher value of the UV excess results from a lower opacity due to metals, and ELS concluded, from this plot, that metal-poor stars – by inference, old stars – formed at all heights above the plane, including distances comparable to halo globular clusters, while metal-rich – young – stars are confined to the plane. Their interpretation was that the oldest stars (i.e., metal-poor, high-velocity, Population II, halo stars) formed during a short-lived rapid collapse some $\sim 10^{10}$ year ago, lasting less than the current Galactic orbital period of the Sun ($\sim 10^8$ year). A gaseous disk formed subsequently, with angular momentum providing support in the radial direction but with no pressure support to prevent continuing collapse in the vertical direction as the gas radiated energy and cooled.

This remarkable analysis laid the foundation for the very large subsequent body of work that uses analyses of the stellar populations of the Milky Way to infer how the Galaxy formed and evolved and, by extension, derive the physics behind the evolution of disk galaxies in general. The ideas within ELS were indeed transformative and their legacy endures as the basis for the “monolithic collapse” scenario of galaxy formation. However, their sample was sufficiently limited that the results were subject to the vagaries of small-number statistics and imperfectly defined selection functions. Further, the UV excess starts to lose sensitivity for sufficiently low metallicities, below around $1/30 Z_{\odot}$ or $[\text{Fe}/\text{H}] \lesssim -1.5$, and the scale is very nonlinear (e.g., Carney 1979). The large wide-field spectroscopic surveys of the current era mitigate these effects.

Globular clusters are intrinsically luminous tracers of the stellar halo and offer a complementary view. Intrinsic variations, from cluster to cluster, in their color-magnitude diagrams were quickly established – indeed, there were a special American Astronomical Society meeting in 1959 devoted to “Differences Among Globular Clusters.” Sandage and Wallerstein (1960) called attention to the fact that the distribution of stars along the horizontal branch (now referred to as the horizontal branch morphology) of a given cluster depended on the iron abundance of the member stars. Metallicity may then be considered to be the “first parameter” affecting the horizontal branch morphology. Anomalies, whereby a pair of clusters of the same metallicity differed in HB morphology, were noted (Sandage and Wildey 1967) and attributed to helium abundance variations (van den Bergh 1967; Faulkner 1966). The helium abundance cannot be determined from direct spectroscopic analysis, so this would be very difficult to establish from observations. Rood and Iben (1968) demonstrated that an age range of $\sim 2 \times 10^9$ year could provide an alternative explanation for the variation in HB morphology at fixed metallicity. These authors argued that the relevant timescale for halo collapse is the current orbital time in the outer halo. They further suggested that there could be metal-rich “Population I” stars that are older than the most metal-deficient halo stars. These radical proposals have strong resonance in much modern work.

Searle and Zinn (1978) developed a photometric metallicity indicator for stars on the red-giant branch (RGB), allowing more distant systems and larger samples of stars within globular clusters to be analyzed. Searle and Zinn supplemented their sample (based on 177 RGB stars in 19 globular clusters) with data from the literature to give a sample of 44 clusters with “tolerable” distances and metallicities, covering similar distance ranges within the Galaxy as the maximum distances reached by the local stars studied by ELS. A plot of metallicity versus Galactocentric distance showed a clear difference between the “inner halo” and the “outer halo,” with the division being around the solar distance of ~ 8 kpc. The most metal-rich clusters are all found within the inner halo, while the outer halo shows no trend of metallicity with distance. Searle and Zinn then turned to the horizontal branch morphology, and found clear “anomaly” or “second parameter” effects for clusters in the outer halo (but not in the inner halo), in that the relative numbers of stars bluewards and redwards of the RR Lyrae gap did not follow the

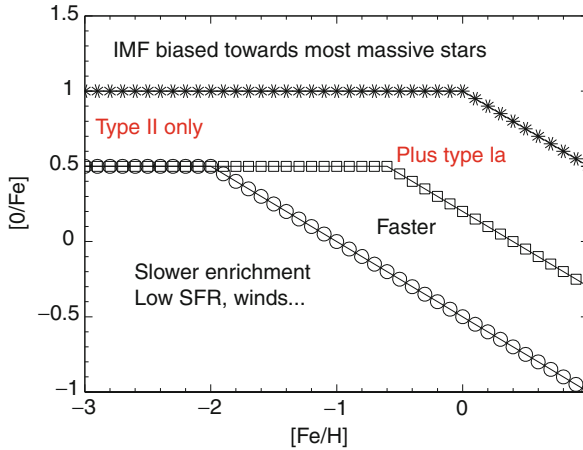
expectations of a simple dependence on metallicity alone (which would lead to a lower fraction of blue HB stars in clusters of higher metallicity). The different behaviors of clusters in the inner and outer regions of the halo led Searle and Zinn to reject the helium abundance as the “second parameter” and to favor age. The large spread in HB morphology observed at a given metallicity, in clusters in the outer halo, led Searle and Zinn to infer an age spread of at least $\sim 10^9$ year, in agreement with the conclusions of Rood and Iben (1968), now extended to a larger sample. Searle and Zinn proposed that their results pointed to a more prolonged period of star and cluster formation in the outer halo than in the inner halo, which showed no evidence of an age range. They further argued that the observed lack of a correlation between metallicity and position implied that the clusters of the outer halo formed in several independent and transient “fragments,” “gaseous condensations,” or “small protogalaxies” before coming into dynamical equilibrium with the overall Galactic potential.

The picture, proposed by Searle and Zinn, of a more extended period of the merging of distinct star- and cluster-forming regions is often contrasted with the rapid “monolithic collapse” of ELS, but it has been realized that aspects of each picture are valid, even if details of each analysis have since been disproven. What we have is a demonstration of the power provided by knowledge of the joint distributions of age, metallicity, kinematics, and spatial distribution for large samples of stars – i.e., the stellar populations – to give insight into how their host galaxies formed and evolved. Larger samples of stars with more detailed, accurate, and precise data have both helped and hindered this insight, as complexities became apparent.

The quantification of the metallicity distributions of stars in the Milky Way of a range of kinematics motivated the development of models of chemical evolution of increasing sophistication (e.g., van den Bergh 1962; Schmidt 1963). Insights from the spectroscopic study of extraGalactic HII regions – “the first metal-poor systems of Population I to be discovered” (Searle and Sargent 1972) – established the “simple closed-box” model as the fundamental basis for comparison with observations. Straightforward extensions to include outflows (Hartwick 1976) and inflows (Mould 1984) demonstrated how a broad range of stellar populations could be created, while maintaining transparent physical assumptions in the models.

The determination of ages of individual field stars – as opposed to cluster members – remains a complex problem. The four-band intermediate-band photometric system introduced by Strömgren (see Strömgren 1987 for a review), supplemented with a measurement of the strength of Balmer-line $H\beta$, remains the most widely used technique. Combinations of the measured magnitudes in these filters have been calibrated for a range of stellar spectral types – in particular F/G dwarfs – to provide estimated values for reddening along the line of sight to the star, metallicity, effective temperature, and gravity. Comparison with theoretical isochrones then allows ages to be determined for even only slightly evolved stars (albeit with large uncertainties). The application of Strömgren photometry has been primarily to local field stars (e.g., Nordstrom et al. 2004; Nissen and Schuster 1991), but the efficacy of this technique for study of members stars in local dwarf galaxies was recently demonstrated by Faria et al. (2007).

Elemental abundance patterns contain a wealth of information beyond that from overall “metallicity,” since different elements are produced in stars of different masses and ejected into the interstellar medium on different timescales (e.g., Tinsley 1979). In particular, the alpha-elements (so-called since they are created by successive addition of a helium nucleus) are synthesized in massive stars, in amounts that depend on the mass of the star, and ejected on short timescales during core-collapse (predominantly Type II) supernovae, while iron has an important contribution from Type Ia supernovae, on much longer timescales, in addition to



■ Fig. 1-3

Schematic pattern of $[O/Fe]$ against $[Fe/H]$ for a self-enriching gas cloud, assuming good mixing and full sampling of the massive-star initial mass function (Adapted from Wyse and Gilmore (1993))

massive stars. The massive-star initial mass function (IMF), the efficiency of mixing and enrichment, together with star-formation timescales all affect the pattern of, e.g., oxygen to iron as a function of iron, as indicated schematically in [Fig. 1-3](#). The observed lack of scatter in the elemental abundance patterns in stars thought to have been (pre-)enriched by only core-collapse supernovae implies there is little variation in the massive-star IMF (e.g., Ruchti et al. 2011). Correlations between the $[\alpha/Fe]$ ratio and kinematics have been noted for decades (e.g., Wallerstein 1962), and modern large-scale surveys are demonstrating the power of this approach, as reviewed by Nissen [Chap. 2](#). The combination of ages, distances, and space motions (often derived from Strömgren photometry) with elemental abundances is particularly powerful. The crucial role played by the detailed elemental abundances of the least enriched stars in constraining the IMF of the first stars – so-called Population III – is discussed in the contribution to this volume by Frebel and Norris [Chap. 3](#). The stellar IMF is reviewed in this volume by Kroupa et al [Chap. 4](#).

3 Some Complexities: What Is Population II?

3.1 Thick Disks

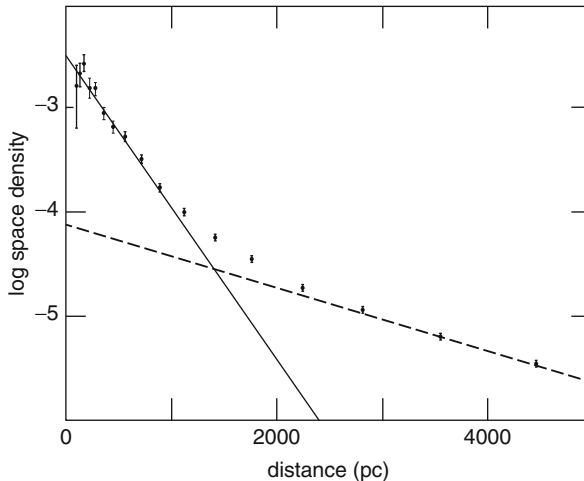
The resolution of stars down to a few magnitudes below the tip of the RGB in the “halo”² of M31, and comparison with fiducials from globular clusters, led to the realization that the typical metallicity of stars in the “halo” of M31 was significantly higher than that of the halo of the Milky Way. Mould and Kristian (1986) found a mean $\langle [M/H] \rangle \sim -0.6$ in a field 7 kpc along the

²More recently, this component has been identified with the bulge of M31 and also with accreted material from a satellite of M31; see [Sect. 3.2.2](#) below.

minor axis of M31, a factor of 10 or so higher than the metallicity of the field halo of the Milky Way, and this value for the dominant population has been confirmed by many subsequent, deeper, studies of stars in fields in the outer parts, over a large fraction of the face of M31 (e.g., Ferguson et al. 2002). This means that “Population II” in the Milky Way and “Population II” in M31, as defined by Baade, cannot represent the same population. Wyse and Gilmore (1988) suggested that “Population II,” as Baade defined it, should refer to the dominant “non-thin-disk” population, and they identified that with the thick disk in the Milky Way, which they had recently found to have a mean metallicity very close to that of the field “halo” in M31 (Gilmore and Wyse 1985), and similar intermediate kinematics (Wyse and Gilmore 1988).

The Galactic thick disk was identified by Gilmore and Reid (1983) as a separate structural component, distinct from the thin disk, through the analysis of counts of F/G stars toward the South Galactic Pole. They derived distances by assuming a smooth metallicity gradient with height that reached $[Fe/H] \sim -0.64$ dex at ~ 2 kpc and adopting an appropriate luminosity function. The resulting stellar space-density distribution was best fit by two (vertical) exponentials, with scale heights of ~ 300 pc and ~ 1 kpc, representing the (old) thin disk and thick disk, respectively (see [Fig. 1-4](#) here). Yoshii (1982) also noted that a single exponential was not a good fit to star counts toward the (North) Galactic Pole, but in his analysis he assumed different stellar metallicity distributions and luminosity functions from those adopted by Gilmore and Reid, and Yoshii described the resulting density distributions in terms of the thin disk and stellar halo. Knowledge of the properties of the stars under study is crucial.

This identification of a distinct Galactic thick disk is reminiscent of the “Intermediate Population II,” introduced as part of the scheme of stellar populations developed in the 1957 Vatican Symposium (Oort 1958; Blaauw 1965). However, as discussed in Wyse and Gilmore (1988), the properties of the proposed Intermediate Population II do not match those of any Galactic stellar population, as we understand them today. The stellar population of the thick disk, sampled



■ Fig. 1-4

The identification of the thick disk of the Milky Way in star counts toward the South Galactic Pole; the derived space density of the stars is best fit by two exponentials (Fig. 6a from Gilmore and Reid (1983), used with permission)

within a few kiloparsec of the Sun (Wyse and Gilmore 1988; Carney et al. 1989; Sandage and Fouts 1987; Gilmore et al. 1995; Nordstrom et al. 2004), has a narrow age range, with the typical star being 10–12 Gyr old, and has a mean iron abundance around one-third of the solar value ($[\text{Fe}/\text{H}] \sim -0.6$). Estimates of the scale length and scale height of the thick disk, together with the local normalization, lead to a stellar mass equal to 10–20% of that of the thin stellar disk (Gilmore et al. 1989; Jurić et al. 2008).

The thick disk, as represented by field stars, has a counterpart in globular clusters, as demonstrated by Zinn (1985): the known globular clusters split into two main components – disk and halo – with distinct kinematics, metallicity distributions, and spatial distributions (with a further division into “inner” and “outer” halo clusters). The properties of the “disk” clusters are very similar to those of the typical (local) thick disk stars. The prototypical (thick) disk cluster is 47 Tuc, which is as old as the more metal-poor halo clusters (e.g., Dotter et al. 2010). The presence of clusters in the thick disk places additional constraints on models of the formation of the thick disk, due to the much greater mass of a cluster compared to that of individual field stars (cf. footnote in Sandage 1981).

Thick disks have been identified in many external galaxies, often associated with bulges – indeed, thick disks were initially detected in the vertical surface brightness profiles of edge-on, high bulge-to-disk ratio S0 galaxies (Tsikoudi 1979; Burstein 1979). Subsequently, thick disks were also detected in later-type spiral galaxies (e.g., van der Kruit and Searle 1981), again through surface photometry. Difficulties in obtaining and calibrating deep, uniform, multiband, wide-field surface photometry hampered the characterization of the stellar populations in the thick disks in external galaxies. More recent detection through star counts, using the Hubble Space Telescope, revealed that these thick disks were also composed of old stars (e.g., Mould 2005; Dalcanton et al. 2007), as in the Milky Way.

Old age was an original defining feature of Population II, and thick disks satisfy that criterion. Their significantly higher mean metallicity than that of the subdwarfs of the Galactic stellar halo is indicative of the many parameters that determine how a cloud of gas turns into stars and indeed how a galaxy evolves.

An in-depth discussion of models for the formation of the thick disk is beyond the scope of this review (see Yoshii’s [▶ Chap. 8](#)). Most models invoke mergers between the young protogalaxy and a companion, either (i) as a means to impart added energy in random motions to the stars of a preexisting thin stellar disk (orbital energy of a merging satellite going into excitations of the internal degrees of freedom of the disk); (ii) to contribute stars, stripped from a merging satellite, directly onto orbits consistent with the thick disk; or (iii) to provide the highly turbulent gas in which early star formation, perhaps in exceptionally massive clusters, produces the thick disk. The old age of stars in the thick disk limits the mergers in models of types (i) and (iii) to have occurred very early. In models of type (ii), if the merger happened more recently than ~ 10 Gyr ago (the age of the stars in the local Galactic thick disk), the accreted satellite(s) cannot have been massive and dense enough to heat the thin disk and must contribute only old stars that match all the properties of the thick disk (this last condition is breached by existing models). We return to the cosmological implications in [▶ Sect. 4](#) below.

3.2 Bulges

The bulge components of disk galaxies share many characteristics with those of elliptical galaxies (e.g., Faber and Gallagher 1979; Kormendy 1985; Bender et al. 1992; Rich [▶ Chap. 6](#))

and have long been considered as smaller analogues of ellipticals. The pioneering modeling of stellar populations³ by Tinsley (1972a, b) and Tinsley and Gunn (1976) showed that the integrated photometric and spectroscopic properties of ellipticals matched the expectations of an old (~10 Gyr) solar-metallicity population, with a short duration of star formation. To first order, the stars in bulges (and in ellipticals) are old and metal rich, again combining the defining characteristics of Population II and Population I. More recently, such bulges are referred to as “classical bulges,” to differentiate them from “pseudobulges,” the nomenclature introduced to categorize those “bulges” with stellar populations that are closer to those of the inner disk of the host galaxy (see Kormendy and Kennicutt 2004 for a review, plus Kormendy and Bender this volume) and may be formed via a disk instability. Both flavors of bulge may coexist (see below).

It is natural that the lowest angular momentum parts of any initially extended gaseous distribution will settle to the central regions, and indeed, models of disk galaxy formation usually identify bulges with that material, whether defined by an initial distribution or after modification by torques (from a range of possible sources – mergers, spiral arms, bar, etc.) after a disk has formed. The properties of the luminous bulge correlate with the mass of the central black hole in galaxies, but intriguingly, the properties of “pseudobulges” do not (Kormendy et al. 2011).

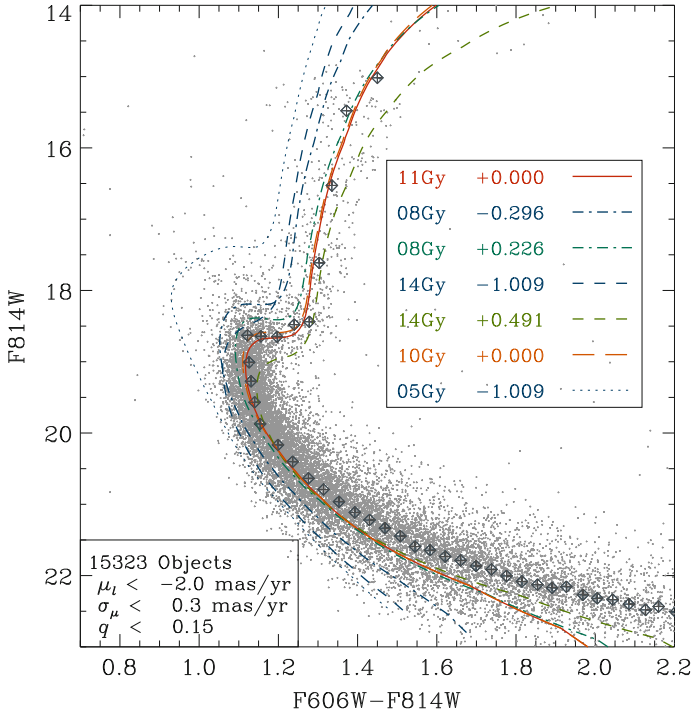
3.2.1 Milky Way Galaxy

The stellar population of the Galactic bulge has been studied most at optical wavelengths (where differences among the properties of the stars are most easily quantified), utilizing the low-reddening lines of sight that were identified by Baade (1951). The most studied of these is “Baade’s Window,” at a projected distance of ~500 pc from the Galactic Center, just 1° away from the minor axis. Whitford (1985) describes how it was immediately apparent from the pioneering spectroscopic work of Morgan, and of himself, that the K and M giants of Baade’s Window were of a mean metallicity equal to that of Population I in the solar neighborhood. Given that Baade had, in his Window, earlier discovered RR Lyrae variables – an evolutionary phase of old, metal-poor stars – a broad range in metallicity was clearly present. This was confirmed by the spectroscopic study of K giants by Rich (1988), and it was soon established that the mean iron abundance of stars in Baade’s window, at the distance of the Galactic Center, was roughly the solar value, with a distribution that was well described by the simple, closed-box model.

Line-of-sight (radial) velocities and proper motions of K-giant stars in Baade’s window were consistent with an isotropic velocity dispersion tensor, and spectroscopic survey fields away from the minor axis, again at projected Galactocentric distances of more than 500 pc, showed a modest amount of net rotation (Ibata and Gilmore 1995): the Milky Way bulge appeared consistent with being moderately flattened by rotation, as found for external bulges and low-luminosity ellipticals (see review of Wyse 1999).

Color-magnitude diagrams of stars toward the bulge are inevitably contaminated by foreground stars, and while an old mean age for stars in the bulge was favored by many investigations, this was predicated on the majority of younger stars in the field being foreground. The exquisite images from the Hubble Space Telescope provided the necessary data for robust foreground subtraction by use of individual stellar proper motions. The resultant “cleaned” color-magnitude diagram for stars in the bulge (based on proper motion) in a field at projected

³The primary motivation was to understand how the evolution of the stellar content of galaxies affected the use of galaxy counts to infer cosmological parameters.

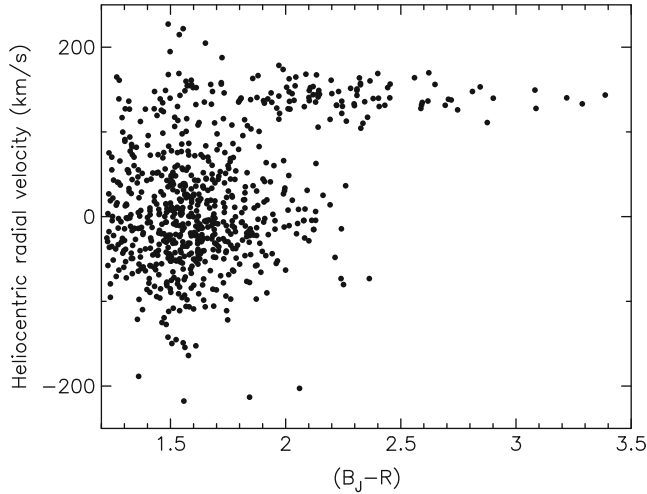


■ Fig. 1-5

Color-magnitude diagram for stars in the Galactic bulge at line of sight at $(\ell, b) = (2.65^\circ, -1.25^\circ)$. Foreground stars have been removed using a threshold in proper motion. The age-metallicity degeneracy inherent in optical colors is evident from the parameters of the theoretical isochrones (age and $[\text{Fe}/\text{H}]$ as given on the figure), but the dominant population is clearly old and metal rich, with a range of metallicities (Fig. 20 from Clarkson et al. (2008), used with permission)

Galactocentric distance of ~ 350 pc is shown in ● Fig. 1-5, taken from Clarkson et al. (2008). The dominant population is clearly old and metal rich (compare with the theoretical isochrones).

Closer to the plane than Baade’s Window (at ~ 500 pc), there is strong photometric evidence for a bar (e.g., Binney et al. 1997) and increasing kinematic evidence (e.g., Shen et al. 2010). How this mildly triaxial structure formed – from an instability in the inner disk? – and how it relates to the more extended oblate, rotationally supported “bulge” are open questions. Study of the elemental abundance distributions as a function of latitude, over the range $b = -4^\circ$ to $b = -12^\circ$, suggests a dual nature, with the most metal-rich stars, found predominantly at low latitudes, associated with the bar/thin disk, and the less metal-rich stars ($[\text{Fe}/\text{H}] < -0.5$), which show similarity to stars in the local thick disk, tracing the “bulge” (Gonzalez et al. 2011). The unevolved stars in the bulge region are generally too faint for spectroscopic work, could they be distinguished from the foreground disk population. However, fortuitous microlensing can cause sufficient brightening of dwarf stars in the bulge that high-resolution spectra can be obtained. Elemental abundances of such magnified dwarf stars show a similar dual nature, either more disklike (a “pseudo” bulge) or more “true” bulge, with the additional dimension of age estimates from Strömgren photometry (e.g., Bensby et al. 2010). The connection between thick



■ Fig. 1-6

The discovery of the Sagittarius dwarf spheroidal galaxy: the stars with a narrow range of velocities around 150 km/s and very red colors are members of the Sgr dwarf galaxy (Fig. 4 from Ibata et al. (1995), used with permission)

disks and bulges mentioned in [Sect. 3.1](#) above appears to be gaining support. The similarity between aspects of the stellar populations of thick disk and bulge indeed suggests a unified formation scenario (e.g., Jones and Wyse 1983; Wyse 2001), and models invoking an early gas-rich turbulent disk have been proposed (e.g., Bournaud et al. 2009). Very large surveys of elemental abundances are within the capabilities of planned surveys and should establish the connections, or otherwise, between and among the inner thin disk, thick disk, and bulge.

It should be noted that the Ibata and Gilmore (1995) wide-area survey of the bulge provided an unexpected result: the discovery of the Sagittarius dwarf spheroidal galaxy. The existence of this distinct satellite galaxy was revealed through its different stellar population, as illustrated in [Fig. 1-6](#), taken from Ibata et al. (1995). This discovery transformed our view of the stellar halo and provides unprecedented insight into the merging/accretion process. As we briefly discuss below, the identification and characterization of satellite galaxies is a very active field.

3.2.2 M31

The resolved stellar population of the bulge in M31 can be studied in the optical, from the ground, only in regions far from the central parts (where crowding is not important) and is limited to intrinsically bright stars. As discussed above ([Sect. 3.1](#)), both spectroscopic and photometric studies of the red-giant branch stars in several small areas distributed across the face of M31 at distances of greater than ~ 10 kpc from the center (mostly along the minor axis to minimize the contribution from M31's disk) were consistent with an old population of mean metallicity around one-third of the solar value. Very wide-field imaging revealed, however, the limitation of such “pencil-beam” surveys: counts of red-giant stars showed very significant variation across the face of M31, as did the characteristic broadband color of these stars, which is the

standard photometric indicator of mean metallicity of the (old) population traced by the RGB (Ferguson et al. 2002). Many of the fields previously believed to be probing the “bulge,” particularly those on the minor axis, were revealed to be substantially contaminated by substructure. The nature of this substructure remains unclear, but it is likely to have been created by a minor merger – tidal debris from the accreted/disrupted satellite galaxy and/or stellar material that was originally in the disk and has been disturbed by the interaction. Similarly to the situation in the Milky Way halo (► Sect. 3.3.1 below), much of the substructure can be ascribed to one satellite progenitor, plus distorted/heated thin disk material (Richardson et al. 2008).

The oldest main-sequence turnoff is within reach of the *Hubble Space Telescope*, and deep imaging in several fields in the “bulge/halo” of M31 (see Brown et al. 2008 and references therein) confirmed the heterogeneous nature of the star-formation histories and chemical enrichment of the stars in each line of sight.

Study of resolved stars in the infrared, especially with the *Hubble Space Telescope*, can probe closer to the center of M31; data for stars within a projected distance of ~ 5 kpc (Stephens et al. 2003) revealed a luminosity function consistent with that of the stars in Baade’s Window. Analysis of the line strengths of spectra obtained through long-slit spectroscopy in the optical over the inner regions of M31 (within a projected distance of a few kiloparsec) implies an old, metal-rich population (Saglia et al. 2010), similar to that of the “classical” bulge of the Milky Way.

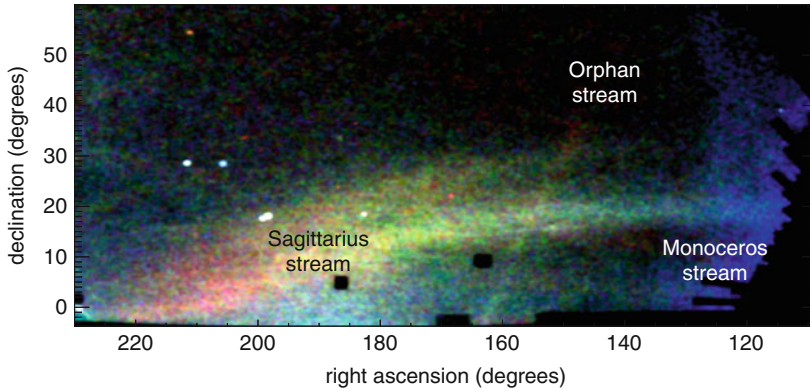
The identification and global definitions of “halo,” “bulge,” and “thick disk” of M31 are ongoing projects. Indeed, the photometric surveys are far ahead, in both depth and area, of the spectroscopic surveys, which are really required to break age/metallicity degeneracies. Planned capabilities for very high multiplexing multi-object spectroscopy on 8–10-m class telescopes in the near future should provide unprecedented data and hopefully understanding too. In the meantime, wide-area photometric surveys are identifying numerous satellite galaxies and globular clusters of M31 that were previously unknown, and we return to the cosmological implications below.

3.3 Stellar Halo

The acquisition, and creation of the final dataset, of the very wide area, uniform precise, and accurate multiband photometry from the Sloan Digital Sky Survey revolutionized study of the stellar populations away from the plane of the thin disk. The now-iconic image of the “Field of Streams” (Belokurov et al. 2006a), reproduced here in ► Fig. 1-7, shows the distribution of old, metal-poor, main-sequence turnoff stars (color- and magnitude-selected) over a significant fraction of the sky, color-coded to represent distance, blue through green to red, with blue representing stars at distances of ~ 10 kpc and red stars at ~ 30 kpc.

3.3.1 Field Stars

It is immediately clear from ► Fig. 1-7 that the distribution of faint stars is far from uniform on the sky. The dominant feature is due to tidal debris from the Sagittarius dwarf spheroidal, the core of which was first detected on the other half of the sky (or Galaxy). The origins of the Orphan Stream, the Monoceros Stream, and, indeed, what appears to be the bifurcated structure of the Sagittarius Stream are all topics of current research.



■ Fig. 1-7

The “Field of Streams” and dots. This shows the distribution of old main-sequence turnoff stars in the Galaxy, as revealed by the photometric data of the Sloan Digital Sky Survey. Courtesy Vasily Belokurov

Quantification of the amount of substructure depends somewhat on the tracer used (Bell et al. 2008; Deason et al. 2011), but the conclusion that the Sagittarius stream plus an additional structure in the direction of Virgo (the “Virgo Overdensity”) are the two dominant sources of deviations from a smooth density law is robust. It might also be noted that these structures are in the *outer* stellar halo, at Galactocentric distances of ~ 15 kpc to ~ 40 kpc, and even over this range, a smooth underlying component can be easily traced. The inner stellar halo, which contains most of the mass, is smooth (as may be expected from the shorter dynamical timescales).


The dominant population of the stellar halo is old, ~ 10 – 12 Gyr (Unavane et al. 1996), the same as for the dominant populations in the (local) thick disk and bulge. The excellent photometry and spectroscopy from the SDSS is consistent with a very uniform age, with no dependence on metallicity (Jofre and Weiss 2011). A non-negligible population with bluer colors than the dominant turnoff, at given metallicity, is also detected; whether these are blue stragglers (descended from binary systems), as appears to be the case in the bulge (Clarkson et al. 2011), or genuinely younger stars remains unclear. Unavane et al. (1996) discuss the implications for the allowable accretion of stars from typical satellite galaxies: this has to have happened very early or involve only a very small fraction of the field halo.

3.4 Satellite Galaxies

The “dots” in ► Fig. 1-7 are also extremely interesting; many of them represent newly discovered satellite systems (Belokurov et al. 2006b), the faintest of which extend down to the regime of very bright individual stars, $\sim 10^3 L_{\odot}$ (Belokurov et al. 2007). Establishing whether these are galaxies – i.e., have a dark-matter halo – or star clusters requires spectroscopy to determine the range in velocities of member stars, and hence the gravitating mass, and also to estimate the internal metallicity spread. The contributions to this volume by Frebel and Norris (► Chap. 3) and Walker (► Chap. 20) provides a discussion of the importance of the dwarf galaxies in understanding the process of star formation at very early epochs, and in quantifying the distribution of dark matter within galaxies, and hence constraining the nature of dark matter.

3.4.1 Globular Clusters

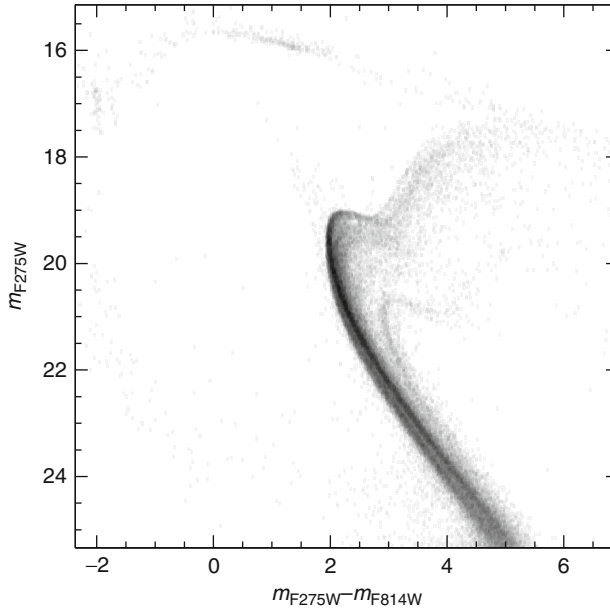
Searle and Zinn gained their intuition into the relative ages of globular clusters from the variation in the clusters' horizontal branch morphologies. The main-sequence turnoff is a more unambiguous measure of age, and again, the excellent image quality of the Hubble Space Telescope provided the means by which this could be determined for a statistical sample of Galactic globular clusters. Analysis of the main-sequence turnoff colors, as a function of metallicity, revealed that age is indeed the best “second parameter” (with metallicity the first parameter) that determines the HB morphology (Dotter et al. 2010). Further, most Galactic globular clusters are old, ~ 12 Gyr, with little variation in age across the entire metallicity range ($+/-1$ Gyr). A small subset of clusters, essentially either in the outer Galaxy (Galactocentric distances of greater than ~ 10 kpc) or associated with the Sgr dSph, are significantly younger. More curiously, they follow an age-metallicity relation (see Fig. 9 of Dotter et al. 2010), suggestive of one self-enriching progenitor rather than late accretion of several systems. The fate of globular clusters in mergers/accretion is an interesting question, as is the reason behind the very high number of globular clusters, relative to host galaxy luminosity, found in dwarf galaxies.

Perhaps the most unexpected result from the Hubble Space Telescope is the presence of multiple populations in globular clusters hitherto believed to be excellent examples of a “Simple Stellar Population,” characterized by a single age and single metallicity. The situation is well illustrated by the bewildering complexity of the very luminous cluster ω Centauri, which had long been known to have an internal spread in metallicity, but was still rather “simple.” Precise proper motions from HST images can be used to remove contamination by field stars in the cluster color-magnitude diagram. The result for ω Cen is illustrated in  Fig. 1-8, taken from Bellini et al. (2010): the clear presence of several main-sequence turnoffs, subgiant branches, and red-giant branches. There is at present no good understanding of how the stars formed and the cluster (self-)enriched. One major puzzle is that the helium content apparently must vary substantially between the subpopulations, with no obvious source for the additional helium.

4 Cosmological Implications of the Properties of (Galactic) Stellar Populations

The now-standard paradigm of structure formation by hierarchical clustering had been developed around the same time that we were gaining the understanding of Galactic stellar populations discussed here. White and Rees (1978) set the scene by investigating the dissipational formation of galaxies within dark-matter haloes, with an assumed power spectrum of primordial density fluctuations such that small scales, subgalactic mass, collapsed first, with larger systems built up by hierarchical clustering and merging. The dissipation of baryons leads to galaxies occupying only the central parts of dark haloes, enhancing their ability to survive the merging process. The comprehensive review by Blumenthal et al. (1984) of galaxy formation with cold dark matter (consisting of very massive, weakly interacting particles) demonstrated the broad agreement with observations over a wide range of physical scales and established this scenario (with the later modification to include dark energy) as the default assumption.

The “concordance” cosmological model, with $\Omega_{\Lambda} \sim 0.73$, $\Omega_{\text{CDM}} \sim 0.23$, and $\Omega_{\text{baryons}} \sim 0.04$, provides an excellent fit to observations of large-scale structure such as galaxy clustering and the



■ Fig. 1-8

Hess diagram, i.e., a plot of the relative density of stars in the apparent magnitude – color plane, for the globular cluster ω Cen, based on data from the Wide Field camera 3 on the Hubble Space Telescope. The unprecedented precision of the data reveals the complexity of the multiple stellar populations present in this cluster (Fig. 10 from Bellini et al. (2010), used with permission)

fluctuations in the microwave background (e.g., Spergel et al. 2007). Smaller scales, however, are where the predictions of different flavors of dark matter diverge (e.g., Ostriker and Steinhardt 2003), and indeed, the Galactic scale is where “challenges” to Λ CDM have recently become apparent. There have been two approaches to these challenges: maintain the dark matter as “cold” and attempt to modify the baryonic physics or explore models with different dark-matter particles.

Many of these challenges stem from the large abundance of persistent small-scale structure in Λ CDM, and the merging inherent in the buildup of larger systems, such as galaxies like the Milky Way. Orders of magnitude more dark sub-haloes than observed satellite galaxies are predicted (Klypin et al. 1999; Moore et al. 1999), and a galaxy like the Milky Way should, within the last 10 Gyr or so, have had a merger with a satellite of mass at least equal to that of the present-day disk (Stewart et al. 2008).

The characterization of the stellar population of the (local) Galactic thick disk as being predominantly old has major implications for the merger history of the Milky Way and for the other disk galaxies with old thick disks. A significant merger that occurred, say, 2 Gyr ago should have heated the thin disk at that time, resulting in a substantial population of 2-Gyr-old stars in the thick disk. The fact that stars younger than ~ 10 Gyr are not present in large numbers (rare runaway events do happen) rules out a “typical” merging history as predicted by Λ CDM.

Mergers also build up bulges, and again, the stellar population in the bulge implies a quiet past for the Milky Way. The identification of many “pseudo-bulges” which may have formed

from an instability in the disk, independently of mergers (see Kormendy and Kennicutt 2004), further diminishes the possible role of mergers, in both the Milky Way and external disk galaxies.

The existence of old stars in the local thin disk provides a similar constraint in that if these stars have remained in their birthplace, this implies the plane of the disk was established a long time ago and that virial equilibrium has not been substantially perturbed in the meantime. Further, star formation was apparently initiated rather early, even several disk scale-lengths from the center. Recent suggestions that stars can “migrate” significant distances while maintaining close to circular orbits (see Sellwood’s contribution ▶ Chap. 18) are stimulating much research.

The attempts at reconciliation between the predicted population of dark satellite systems with the observed satellite galaxies of the Milky Way and of M31 have had some success as far as simple numbers are concerned, by invoking efficient “feedback” to suppress star formation, but many aspects such as the luminosity function, the spatial distribution, and indeed the stellar populations remain open issues. The field halo of the Milky Way, to be produced by disrupted substructure, is rather uniform in its properties, with the exception of the substructure in the outer parts noted above.

Merging is ongoing in both the Milky Way and M31 but is not dominant now and arguably never was. The properties of stellar populations on galactic scales highlight the tensions between the predictions of the “standard” cosmological scenario and observations. We indeed live in interesting times.

Acknowledgments

I acknowledge support through grants AST-0908326 and CDI-1124403 from the National Science Foundation. I thank the Aspen Center for Physics, supported by NSF Grant 1066293, for hospitality while I completed this chapter.

Cross-References

- ▶ [Chemical Abundances as Population Tracers](#)
- ▶ [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- ▶ [Dynamics of Disks and Warps](#)
- ▶ [Galactic Distance Scales](#)
- ▶ [Globular Cluster Dynamical Evolution](#)
- ▶ [High-Velocity Clouds](#)
- ▶ [History of Dark Matter in Galaxies](#)
- ▶ [Interstellar PAHs and Dust](#)
- ▶ [Mass Distribution and Rotation Curve in the Galaxy](#)
- ▶ [Metal-Poor Stars and the Chemical Enrichment of the Universe](#)
- ▶ [Open Clusters and their Role in the Galaxy](#)
- ▶ [Star Counts and the Nature of the Galactic Thick Disk](#)
- ▶ [The Galactic Bulge](#)
- ▶ [The Infrared Galaxy](#)
- ▶ [The Stellar and Sub-Stellar Initial Mass Function of Simple and Composite Populations](#)

References

- Baade, W. 1944, *ApJ*, 100, 137
- Baade, W. 1951, *Publ. Obs. Univ. Michigan*, 10, 7
- Bell, E., et al. 2008, *ApJ*, 680, 295
- Bellini, A., Bedin, L. R., Piotto, G., Milone, A., Marino, A., & Villanova, S. 2010, *AJ*, 140, 631
- Belokurov, V., et al. 2006a, *ApJ*, 642, L137
- Belokurov, V., et al. 2006b, *ApJ*, 647, L111
- Belokurov, V., et al. 2007, *ApJ*, 654, 897
- Bender, R., Burstein, D., & Faber, S. M. 1992, *ApJ*, 399, 462
- Bensby, T., et al. 2010, *A&A*, 512, 41
- Binney, J., Gerhard, O., & Spergel, D. 1997, *MNRAS*, 288, 365
- Blaauw, A. 1965, in *Galactic Structure, Stars and Stellar Systems*, Vol. 5, ed. A. Blaauw, M. Schmidt (Chicago: University of Chicago Press), 435
- Blumenthal, G. R., Faber, S. M., Primack, J. R., & Rees, M. J. 1984, *Nature*, 311, 517
- Bournaud, F., Elmegreen, B., & Martig, M. 2009, *ApJ*, 707, L1
- Burstein, D. 1979, *ApJ*, 234, 829
- Brown, T. M., et al. 2008, *ApJ*, 685, L121
- Carney, B. W. 1979, *ApJ*, 233, 211
- Carney, B. W., Latham, D. W., & Laird, J. B. 1989, *AJ*, 97, 423
- Clarkson, W., et al. 2008, *ApJ*, 684, 1110
- Clarkson, W., et al. 2011, *ApJ*, 735, 37
- Dalcanton, J. J., Seth, A. C., & Yoachim, P. 2007, in *Island Universes, Astrophysics and Space Science Proceedings* (New York: Springer, p. 29) (arXiv:astro-ph/0509700)
- Deason, A. J., Belokurov, V., & Evans, N. W. 2011, *MNRAS*, 416, 2903
- Dotter, A., et al. 2010, *ApJ*, 708, 698
- Eggen, O., Lynden-Bell, D., & Sandage, A. R. 1962, *ApJ*, 136, 748 (ELS)
- Faber, S. M., & Gallagher, J. S. 1979, *ARAA*, 17, 135
- Faria, D., et al. 2007, *A&A*, 465, 357
- Faulkner, J. 1966, *ApJ*, 144, 978
- Ferguson, A. M., et al. 2002, *AJ*, 124, 1452
- Gilmore, G., & Reid, I. N. 1983, *MNRAS*, 202, 1025
- Gilmore, G., & Wyse, R. F. G. 1985, *AJ*, 90, 2015
- Gilmore, G., Wyse, R. F. G., & Jones, J. B. 1995, *AJ*, 109, 1095
- Gilmore, G., Wyse, R. F. G., & Kuijken, K. 1989, *ARAA*, 27, 555
- Gonzalez, O. A., et al. 2011, *A&A*, 530, 54
- Hartwick, F. D. A. 1976, *ApJ*, 209, 418
- Ibata, R., & Gilmore, G. 1995, *MNRAS*, 275, 605
- Ibata, R., Gilmore, G., & Irwin, M. 1995, 277, 871
- Jofre, P., & Weiss, A. 2011, *A&A*, 533, 59
- Jones, B. J. T., & Wyse, R. F. G. 1983, *A&A*, 120, 165
- Jurić, M., et al. 2008, *ApJ*, 673, 864
- Kormendy, J. 1985, *ApJ*, 295, 73
- Klypin, A., Kravtsov, A., Valenzuela, O., & Prada, F. 1999, *ApJ*, 522, 82
- Kormendy, J., & Kennicutt, R. C. 2004, *ARAA*, 42, 603
- Kormendy, J., Bender, R., & Cornell, M. E. 2011, *Nature*, 469, 374
- Moore, B., et al. 1999, *ApJ Letters*, 524, L19
- Mould, J. R. 1984, *PASP*, 96, 773
- Mould, J. R. 2005, *AJ*, 129, 698
- Mould, J. R., & Kristian, J. 1986, *ApJ*, 305, 591
- Nissen, P., & Schuster, W. 1991, *A&A*, 251, 457
- Nordström, B. et al. 2004, *A&A*, 418, 989
- Oort, J. H. 1926, *Pub. Kapteyn Astron. Lab.*, 40, 1
- Oort, J. H. 1958, in *Ricerche Astronomiche*, Vol. 5, Proceedings of a Conference at Vatican Observatory, May 1957, ed. D. J. K. O'Connell (Amsterdam: North-Holland/New York: Interscience), 415
- Ostriker, J. P., & Steinhardt, P. 2003, *Science* 300, 1909
- Rich, R. M. 1988, *AJ*, 95, 828
- Richardson, J., et al. 2008, *AJ*, 135, 1998
- Roman, N. G. 1955, *ApJS*, 2, 195
- Ruchti, G., et al. 2011, *ApJ*, 737, 9
- Rood, R., & Iben, I. 1968, *ApJ*, 154, 215
- Saglia, R. P., Fabricius, M., Bender, R., Montalto, M., Lee, C.-H., Riffeser, A., Seitz, S., Morganti, L., Gerhard, O., & Hopp, U. 2010, *A&A*, 509, 61
- Sandage, A. R. 1969, *ApJ*, 158, 1115
- Sandage, A. R. 1981, *AJ*, 86, 1643
- Sandage, A. R., & Eggen, O. 1959, *MNRAS*, 119, 278
- Sandage, A. R., & Fouts, G. 1987, *AJ*, 93, 592
- Sandage, A. R., & Wallerstein, G. 1960, *ApJ*, 131, 598
- Sandage, A. R., & Wildey, R. 1967, *ApJ*, 150, 469
- Schmidt, M. 1963, *AJ*, 137, 758
- Schwarzschild, M., Searle, L., & Howard, R. 1955, *ApJ*, 122, 353
- Searle, L., & Sargent, W. L. W. 1972, *ApJ*, 173, 25
- Searle, L., & Zinn, R. 1978, *ApJ*, 225, 357
- Shen, J., Rich, R. M., Kormendy, J., Howard, C., De Propris, R., & Kunder, A. 2010, *ApJ*, 720, L72
- Spergel, D., et al. 2007, *ApJS*, 170, 377
- Stephens, A. W., Frogel, J. A., DePoy, D. L., Freedman, W., Gallart, C., Jablonka, P., Renzini, A., Rich, R. M., & Davies, R. 2003, *AJ*, 125, 2473
- Stewart, K., et al. 2008, *ApJ*, 683, 597
- Strömgren, B., 1987, in *The Galaxy, Proceedings of the NATO Advanced Study Institute*, Cambridge, ed. G. Gilmore, & B. Carswell (Dordrecht: D. Reidel), 229
- Tinsley, B. 1972a, *A&A*, 20, 383
- Tinsley, B. 1972b, *ApJ*, 178, 319
- Tinsley, B. 1979, *ApJ*, 229, 1046

- Tinsley, B., & Gunn, J. 1976, *ApJ*, 203, 52
- Tsikoudi, V. 1979, *ApJ*, 234, 842
- Unavane, M., Wyse, R. F. G., & Gilmore, G. 1996, *MNRAS*, 278, 727
- van den Bergh, S. 1962, *AJ*, 67, 486
- van den Bergh, S. 1967, *PASP*, 79, 460
- van der Kruit, P., & Searle, L. 1981, *A&A*, 95, 105
- Wallerstein, G. 1962, *ApJS*, 6, 407
- White, S. D. M., & Rees, M. J. 1978, *MNRAS*, 183, 341
- Whitford, A. 1985, *PASP*, 97, 205
- Wyse, R. F. G. 1999, in *The Formation of Galactic Bulges*, Cambridge Contemporary Astrophysics, ed. C. M. Carollo, H. C. Ferguson, & R. F. G. Wyse (Cambridge, UK/New York: Cambridge University Press), 195 (arXiv:astro-ph/0003150)
- Wyse, R. F. G. 2001, in *Galaxy Disks and Disk Galaxies*, ASP Conference Series, Vol. 230, ed. José G. Funes, S. J., E. M. Corsini (San Francisco: ASP), 71 (arXiv:astro-ph/0012270)
- Wyse, R. F. G., & Gilmore, G. 1988, *AJ*, 95, 1404
- Wyse, R. F. G., & Gilmore, G. 1993, in *The Globular Clusters-Galaxy Connection*, ASP Conference Series, Vol. 48, ed. G. H. Smith, & J. P. Brodie (San Francisco: ASP), 727
- Yoshii, Y. 1982, *PASJ*, 34, 365
- Zinn, R. 1985, *ApJ*, 293, 424

2 Chemical Abundances as Population Tracers

Poul Erik Nissen

Department of Physics and Astronomy, University of Aarhus,
Aarhus C, Denmark

1	<i>Introduction</i>	23
2	<i>Determination of Stellar Abundance Ratios</i>	24
2.1	Observation and Reduction of Stellar Spectra	24
2.2	Model Atmospheres	26
2.3	Abundance Analysis	26
2.4	Determination of Atmospheric Parameters for F, G, and K Stars	28
2.5	Diffusion and Dust-Gas Separation of Elements	28
3	<i>Elements Used as Stellar Population Tracers</i>	29
3.1	Carbon and Oxygen	29
3.2	Intermediate-Mass Elements	31
3.3	The Iron-Peak Elements	32
3.4	The Neutron Capture Elements	34
4	<i>The Galactic Disk</i>	35
4.1	The Thick and The Thin Disk	35
4.2	The $[\alpha/\text{Fe}]$ Distribution of Disk Stars	37
4.3	Abundance Gradients in the Disk	41
5	<i>The Galactic Bulge</i>	41
6	<i>The Galactic Halo</i>	43
6.1	Evidence of Two Distinct Halo Populations	43
6.2	Kinematics and Origin of the Two Halo Populations	45
6.3	Globular Clusters and Dwarf Galaxies	48
7	<i>Conclusions</i>	51
	<i>References</i>	53

Abstract: A discussion of elemental abundance ratios as tracers of stellar populations is presented. The emphasis is on F, G, and K stars because they represent a wide range of ages and have atmospheres providing a “fossil” record of the chemical evolution of the Galaxy.

Instrumentation and methods to determine chemical abundances in stellar atmospheres are discussed in ► Sect. 2. High-resolution ($R > 20,000$) spectra are required to derive precise abundance ratios, but lower resolution spectra may be useful in connection with large statistical studies of populations. Most abundance analyses are based on homogeneous 1D model atmospheres and the assumption of local thermodynamic equilibrium (LTE), but recent works have shown that 3D non-LTE corrections can change the derived trends of abundance ratios as a function of stellar metallicity significantly. However, when comparing stars having similar effective temperatures, surface gravities, and metallicities, 3D non-LTE corrections tend to cancel out. Such a differential approach is the best way to disentangle stellar populations on the basis of chemical abundances.

Abundance ratios particularly useful as population tracers are discussed in ► Sect. 3, including C/O, Na/Fe, Ni/Fe, Ba/Y, Eu/Ba, and α /Fe, where α is the average abundance of Mg, Si, Ca, and Ti. The nucleosynthesis of the elements involved occurs on different timescales in stars and supernovae with different masses. This is the main reason that these abundance ratios can be used as population tracers.

The following sections deal with a discussion of populations in the Galactic disk, the bulge, and the halo. Based on abundance ratios, there is clear evidence for two main populations in the disk: an old, thick disk formed on a timescale of $\sim 10^9$ years and a younger, thin disk formed over a more extended period. For the bulge, interesting new abundance results have been obtained in recent years, including data from microlensed dwarfs, but it is too early to draw any robust conclusions about how and when the bulge formed. For the halo, there is evidence for the existence of two discrete populations with low and high values of α /Fe, respectively. The “low- α ” population has probably been accreted from dwarf galaxies, whereas the “high- α ” population may consist of ancient disk stars “heated” to halo kinematics by merging satellite galaxies. Globular clusters stand out from the halo field stars by showing Na-O and Al-Mg anticorrelations; there is increasing evidence that they consist of multiple stellar populations.

Keywords: Galaxy: bulge, Galaxy: disk, Galaxy: halo, Galaxies: dwarf, Globular clusters: general, Stars: abundances, Stars: atmospheres, Techniques: spectroscopic

List of Abbreviations: *AGB*, Asymptotic giant branch; *CDM*, Cold dark matter; *CRIRES*, Cryogenic infrared echelle spectrograph; *DEIMOS*, DEep Imaging Multi-Object Spectrograph; *DLA*, Damped Lyman-alpha; *dSph*, dwarf spheroidal; *ESA*, European Space Agency; *ESO*, European Southern Observatory; *EW*, Equivalent width; *FIES*, Fiber-fed Echelle Spectrograph; *FLAMES*, Fiber Large Array Multi-Element Spectrograph; *GAIA*, (ESA mission for exploring the Galaxy); *HARPS*, High Accuracy Radial velocity Planet Searcher; *HFS*, Hyperfine structure; *HIRES*, High-resolution echelle spectrometer; *H-R*, Hertzsprung-Russell; *IMF*, Initial mass function; *LMC*, Large Magellanic Cloud; *LSR*, Local standard of rest; *LTE*, Local thermodynamic equilibrium; *NOT*, Nordic Optical Telescope; *RAVE*, Radial Velocity Experiment; *SDSS*, Sloan Digital Sky Survey; *SNe*, Supernovae; *UVES*, Ultraviolet and Visible Echelle Spectrograph; *VLT*, Very Large Telescope

1 Introduction

A population consists of a group of stars with a common origin and history. Hence, it is of high importance for studies of the formation and evolution of the Galaxy to detect and describe existing Galactic populations. This may be done by analyzing distribution functions for stars in space, kinematics, age, and chemical composition. In particular, it is important to know if the main Galactic components, the disk, the bulge, and the halo, each consists of a single stellar population or if multiple populations are needed to fit data for kinematics, ages, and abundance ratios of stars belonging to these components.

Whereas the original spatial and kinematical distributions of stars in a population are modified during the dynamical evolution of the Galaxy, it is generally assumed that the chemical composition of a stellar atmosphere provides a “fossil” record of the composition of the Galaxy at the time and the place for the formation of the star. In this connection, F and G main-sequence and subgiant stars are of particular interest because they span an age range as long as the lifetime of the Galaxy. Furthermore, they have an upper convection zone that mixes matter in the atmosphere with deeper layers, which tends to reduce abundance changes induced by diffusion or accretion processes (see discussion in [Sect. 2.5](#)). On the other hand, the convection zone is not so deep that elements produced by nuclear reactions in the stellar interior are brought up to the stellar surface. Hence, chemical abundances of F and G main-sequence and subgiant stars are expected to be good tracers of stellar populations.

Stars with spectral types different from F and G are also of importance as tracers of Galactic populations. O, B, and A stars may be used to probe the present composition of the Galaxy, but in some cases, the atmospheric composition is affected by diffusion or accretion processes. K giants are very useful as a supplement to the F and G main-sequence stars because they can be observed to greater distances. Their space density is, however, smaller than that of F and G dwarfs, and care should be taken because the atmospheric abundances of some elements, e.g., C and N, may be affected by convective dredge-up of the products of nuclear processes in the stellar interior.

The present review deals with the use of stellar abundance ratios to disentangle the various stellar populations in the Galaxy. Methods to determine chemical abundances in stellar atmospheres are discussed in [Sect. 2](#) with emphasis on the high precision that may be obtained when analyzing stars in a limited region of the H-R diagram differentially. In [Sect. 2](#), it is also discussed if element abundances in F and G main-sequence stars are affected by diffusion or accretion processes. [Section 3](#) contains an inventory of abundance ratios that are particularly useful as tracers of stellar populations and a discussion of the nucleosynthesis of the involved elements. The following [Sects. 4](#), [5](#), and [6](#) deal with populations associated with the Galactic disk, the bulge, and the halo including globular clusters and satellite galaxies. Relations between abundance ratios, kinematics, and ages will be reviewed, and scenarios for the origin of the various populations will be discussed. Finally, [Sect. 7](#) contains conclusions and some thoughts about future observing programs related to chemical abundances as population tracers.

This chapter focuses on stars with metallicities in the range $-3.0 < [\text{Fe}/\text{H}] < +0.4$, where $[\text{Fe}/\text{H}]$ is a logarithmic measure of the ratio between the number of iron and hydrogen atoms in the star relative to the same ratio in the Sun.¹ Extremely metal-poor stars with $[\text{Fe}/\text{H}] < -3.0$ are discussed by Frebel and Norris ([Chap. 8, Metal-poor Stars and](#)

¹For two elements X and Y, $[X/Y] = \log(N_X/N_Y)_{\text{star}} - \log(N_X/N_Y)_{\text{Sun}}$, where N_X and N_Y are the number densities of the elements.

the [Chemical Enrichment of the Universe](#)). To some extent, their chemical abundances are related to single supernovae (SNe) events, whereas a mixture of SNe with a mass distribution determined by the initial mass function (IMF) have produced the elements in more metal-rich stars.

2 Determination of Stellar Abundance Ratios

2.1 Observation and Reduction of Stellar Spectra

In order to derive precise abundance ratios, high-resolution ($R = \lambda/\Delta\lambda > 30,000$) and high signal-to-noise ($S/N > 100$) spectra should ideally be obtained. For such spectra, it is possible to define a reliable continuum and measure equivalent widths of weak spectral lines that have high sensitivity to abundance changes and low sensitivity to broadening parameters as micro-turbulence and collisional damping. Thanks to the installation of efficient echelle spectrographs in connection with many large- and medium-sized telescopes, a large number of high-quality optical ($3,700 < \lambda < 9,000 \text{ \AA}$) spectra for F, G, and K stars have been obtained during the last couple of decades. The infrared spectral region is still lacking behind, but important abundance results for K giants in the Galactic bulge have been obtained with the Phoenix spectrograph on the Gemini South telescope (e.g., Meléndez et al. 2008) and with the ESO VLT cryogenic infrared echelle spectrograph, CRIRES (Ryde et al. 2010).

Spectra with somewhat lower resolution ($R \sim 20,000$) and $S/N \sim 50$ can also be used for determining abundance ratios and may be obtained with multi-object spectrographs such as FLAMES at the ESO VLT. This has proven to be a very effective way of getting abundance data for stars in globular clusters (e.g., Carretta et al. 2009) and satellite galaxies (see review of Tolstoy et al. 2009). Furthermore, abundances of elements that are represented by many lines in stellar spectra, such as Fe and the α -capture elements Mg, Si, Ca, and Ti, can be obtained from medium-resolution spectra ($R \sim 5,000\text{--}10,000$). A good example is the determination of stellar abundances for the Sculptor dwarf spheroidal (dSph) galaxy by Kirby et al. (2009) with the multi-object spectrograph, DEIMOS, at the Keck II telescope.

Even low-resolution ($R \sim 2,000$) spectra are useful for statistical investigations of $[\alpha/\text{Fe}]^2$ in Galactic surveys such as the Sloan Digital Sky Survey (SDSS) (Lee et al. 2011). Another large survey, the Radial Velocity Experiment (RAVE), which will deliver medium-resolution spectra ($R \simeq 7,500$) of $\sim 10^6$ stars in the near-infrared Ca II-triplet region (8,410–8,795 Å), has also the potential of supplying $[\alpha/\text{Fe}]$ with a decent precision (Boeche et al. 2008). In the future, the ESA GAIA mission will make it possible to determine $[\alpha/\text{Fe}]$ values for a still larger sample of stars based also on spectra in the Ca II-triplet region, but with a somewhat higher resolution, $R \simeq 11,500$.

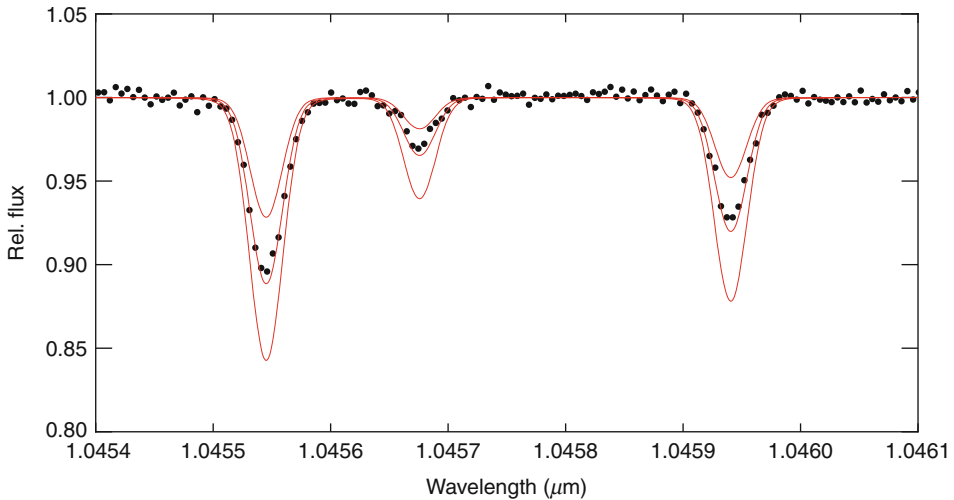
The reduction of raw spectral data should include background and sky subtraction, flat-field correction, extraction of spectra, and wavelength calibration. Standard IRAF³ tasks or special software can be used. Care should be taken to perform a good flatfielding, including removal of possible interference fringes such that a reliable continuum can be defined from wavelength

²Throughout this chapter, α refers to the average abundance of Mg, Si, Ca, and Ti, i.e., $[\alpha/\text{Fe}] = \frac{1}{4} ([\text{Mg}/\text{Fe}] + [\text{Si}/\text{Fe}] + [\text{Ca}/\text{Fe}] + [\text{Ti}/\text{Fe}])$.

³IRAF is distributed by the National Optical Astronomy Observatories, which are operated by the Association of Universities for Research in Astronomy, Inc., under cooperative agreement with the National Science Foundation.

regions free of spectral lines. After normalization of the spectra, equivalent widths (EWs) of weak spectral lines can be measured by Gaussian fitting of the line profiles. For spectral lines having significant line wings (typically $EW > 70 \text{ m\AA}$ in F and G main-sequence stars), the fitting should be performed with a Voigt profile. The continuum setting and equivalent width measurements may be done interactively with the IRAF task `splot` or can be done automatically (e.g., Sousa et al. 2007).

Instead of using equivalent widths, abundances can be determined by fitting a synthetic spectral line profile, calculated for a model atmosphere, to the observed line profile using the abundance of the element as a free parameter of the fit (see [Fig. 2-1](#)). For a line significantly blended by other spectral lines, this is the only way to derive a reliable abundance. Other lines have to be used to determine element abundances for the blending lines and to estimate line broadening parameters associated with stellar rotation and macroturbulence. Hence, the fitting is an iterative process that involves several line regions. It can be done automatically and may include a determination of the basic atmospheric parameters, effective temperature T_{eff} , and surface gravity g , from ratios between selected lines. A good example of an automatic method is presented by Barklem et al. (2005), who determine abundances of 22 elements from ESO VLT/UVES spectra by fitting hundreds of spectral windows containing suitable lines. The procedure includes the identification of continuum points in the windows, adjustments of line centers, rejection of lines disturbed by cosmic-ray hits, and χ -square minimization of the difference between the synthetic and the observed spectrum. For low- and medium-resolution spectra, for which individual lines are not resolved and the continuum is not reached, abundances can be determined by similar methods (e.g., Lee et al. 2011; Kirby et al. 2009) or may be based on line indices that can be calibrated via high-resolution data or by model-atmosphere calculations.



■ Fig. 2-1

The ESO/VLT CRIFRES spectrum of the turnoff halo star G 29-23 ($[\text{Fe}/\text{H}] = -1.7$) around the near-IR, $1.046 \mu\text{m}$ S I triplet (dots) compared with synthetic LTE model-atmosphere profiles for three sulfur abundances, corresponding to $[\text{S}/\text{Fe}] = 0.0, 0.3,$ and 0.6 , respectively. As seen, the $[\text{S}/\text{Fe}] = 0.3$ case provides a close fit to the observations. The average S abundance determined from the three lines corresponds to $[\text{S}/\text{Fe}] = 0.27$ (Nissen et al. 2007)

2.2 Model Atmospheres

Stellar abundances are normally based on a model-atmosphere analysis of the available spectra. In most cases, a plane-parallel, homogeneous (1D) model is adopted, and it is assumed that the distributions of atoms over the possible excitation and ionization states are given by Boltzmann's and Saha's equations. This condition is called "local thermodynamic equilibrium" (LTE). The temperature structure of the model is derived from the requirement that the total flux of energy as transported by radiation and convection should be constant throughout the atmosphere and given by

$$F = \sigma T_{\text{eff}}^4, \quad (2.1)$$

where σ is the Stefan-Boltzmann constant, and T_{eff} is the effective temperature of the star. Furthermore, the atmosphere is assumed to be in hydrostatic equilibrium, and the pressure P as a function of optical depth τ is determined from the equation

$$\frac{dP}{d\tau} = \frac{g}{\kappa_c(T, P_e)}, \quad (2.2)$$

where g is the gravity in the stellar atmosphere and κ_c the continuous absorption coefficient, as determined primarily by H^- absorption in optical and infrared spectra of F, G, and K stars. For these cool stars, electrons in the stellar atmosphere mainly come from the ionization of elements like Mg, Si, and Fe, and the relation between total pressure and electron pressure P_e therefore depends on both metallicity and the α/Fe ratio.

Details of the construction of 1D stellar models may be found in textbooks on stellar atmospheres. The most used grid of models are the ATLAS9 models of Kurucz (1993) and the Uppsala MARCS models (Gustafsson et al. 2008). In both sets of models, convection is treated in the classical mixing-length approximation.

As reviewed by Asplund (2005), 1D models give only a first approximation to the temperature structures of stellar atmospheres. The convection creates an inhomogeneous structure with hot rising granules and cool downflows. Inhomogeneous (3D) models can be constructed by solving the standard equations for conservation of mass, momentum, and energy in connection with the radiative transfer equation for a representative volume of the stellar atmosphere. The mean temperature structure of such 3D models may differ significantly from that of 1D models especially in the case of metal-poor stars. Due to the expansion of rising granulation elements and the lack of radiative heating when the line absorption coefficient is small, the 3D models have much lower temperature and electron pressure in the upper layers than classical 1D models in radiative equilibrium.

2.3 Abundance Analysis

For a given model atmosphere, the flux F_λ in an absorption line can be calculated by solving the transfer equation. Integration over the line profile relative to the continuum flux F_c then gives the equivalent width

$$EW = \int \frac{F_c - F_\lambda}{F_c} d\lambda. \quad (2.3)$$

It is the ratio between the line and continuous absorption coefficients, κ_l/κ_c , that determines the line depth and hence the equivalent width. For a weak (unsaturated) line, the equivalent

width is approximately proportional to the abundance ratio N_X/N_H , where X is the element corresponding to the line. For saturated lines, the equivalent width also depends on line broadening due to small-scale turbulent gas motions. In 1D modeling this introduces an additional atmospheric parameter, the microturbulence, that can be determined from the requirement that the same Fe abundance should be derived from weak and medium-strong Fe I lines. Strong lines with damping wings are sensitive to the value of the collisional damping constant. Clearly, the most accurate abundances are derived from weak lines if observed with high resolution and S/N .

The equivalent width of a line also depends on the oscillator strength and the populations of the energy levels corresponding to the line. In LTE, Boltzmann's and Saha's equations are used to determine the population numbers. This may, however, be a poor approximation as reviewed by Asplund (2005). Instead, one can use that a stellar atmosphere is in a steady state, i.e., that the population n_i of a level i does not vary in time. This can be expressed as

$$n_i \sum_{j=1}^N (R_{ij} + C_{ij}) = \sum_{j=1}^N n_j (R_{ji} + C_{ji}), \quad (2.4)$$

where R and C are the transition rates for radiative and collisional processes, respectively. The summation is extended over all N levels with $j \neq i$. In such, so-called non-LTE calculations, the population numbers are found by solving N equations of the same type as (2.4). In addition, the transfer equation must be solved because the radiative transition rates depend on the mean intensity of the radiation.

Departures from LTE can be large and affect derived stellar abundances very significantly (Asplund 2005). However, in some cases, collisional transition rates are not well known, and the calculated non-LTE populations become rather uncertain. In particular, this is the case for inelastic collisions with neutral hydrogen atoms. Often, the recipes of Drawin (1969) are adopted, but since these estimates are based on classical physics, they only provide an order-of-magnitude estimate of the collisional rates. Hence, a scaling factor S_H to the Drawin formula has to be introduced. It may be calibrated on the basis of solar spectra by requesting that lines with different excitation potential and from different ionization stages should provide the same abundance or one may vary S_H to investigate how the uncertainty of collisional rates affects the derived abundances.

Given that non-LTE calculations are sometimes uncertain and that a grid of 3D models is not yet available, a *differential* 1D LTE analysis is often applied to determine abundance ratios. For narrow ranges in the basic atmospheric parameters, say ± 400 K in T_{eff} , ± 0.4 dex in $\log g$, and ± 0.5 dex in $[\text{Fe}/\text{H}]$, one may assume that non-LTE and 3D effects on the abundances are about the same for all stars. Hence, precise differential abundances with respect to a standard star can be derived in LTE.

For F and G stars with metallicities around $[\text{Fe}/\text{H}] = 0$, the Sun is an obvious choice as a standard, and logarithmic abundance ratios with respect to the Sun, like $[\text{Mg}/\text{Fe}]$, can be derived from the same lines in the spectra of the stars and the Sun. At lower metallicities, bright stars with well-known atmospheric parameters can be chosen as standards. This method has the additional advantage that the oscillator strength of a line cancels out so that its error plays no role. Such differential abundance ratios can be determined to a precision of about ± 0.03 dex (e.g., Neves et al. 2009; Nissen and Schuster 2010). When using chemical abundances to trace stellar populations, it is just these very precise differential abundance ratios at a given metallicity that are needed. Trends of abundance ratios as a function of $[\text{Fe}/\text{H}]$ derived under the LTE assumption are, on the other hand, less accurate, because non-LTE and 3D effects change with metallicity.

2.4 Determination of Atmospheric Parameters for F, G, and K Stars

In order to determine precise abundance ratios, reliable values of the stellar atmospheric parameters, T_{eff} , g , and $[\text{Fe}/\text{H}]$, must be determined. Some abundance ratios like $[\text{Mg}/\text{Fe}]$ determined from neutral atomic lines are fairly insensitive to errors in the atmospheric parameters, but other ratios like $[\text{O}/\text{Fe}]$ with the oxygen abundance determined from the high-excitation O I triplet or from OH lines depend critically on the adopted values for T_{eff} and g .

The effective temperature of a late-type star can be determined from a color index, e.g., $V - K$, calibrated in terms of T_{eff} by the infrared flux method. Two recent implementations of this method (González Hernández and Bonifacio 2009; Casagrande et al. 2010) give consistent calibrations of $V - K$. In the case of nearby stars for which colors are not affected by interstellar reddening, T_{eff} can be determined to an accuracy of the order of ± 50 K. For more distant stars, the reddening is, however, a problem and T_{eff} is better determined spectroscopically, e.g., from the wings of Balmer lines or from the requirement that $[\text{Fe}/\text{H}]$ derived from Fe I lines should be independent of the excitation potential of the lines. In this way, differential values of T_{eff} can be determined to a precision of ± 25 K (Nissen 2008).

The best way to determine the stellar surface gravity

$$g = G \frac{\mathcal{M}}{R^2} \quad (2.5)$$

is to estimate the mass \mathcal{M} from stellar evolutionary tracks and the radius R from the basic relation $L \propto R^2 T_{\text{eff}}^4$, where L is the luminosity of the star. This leads to the following expression for the gravity of a star relative to that of the Sun ($\log g_{\odot} = 4.44$ in the cgs system)

$$\log \frac{g}{g_{\odot}} = \log \frac{\mathcal{M}}{\mathcal{M}_{\odot}} + 4 \log \frac{T_{\text{eff}}}{T_{\text{eff},\odot}} + 0.4(M_{\text{bol}} - M_{\text{bol},\odot}), \quad (2.6)$$

where M_{bol} is the absolute bolometric magnitude, which can be determined from the apparent magnitude if the distance to the star is known.

This method of determining surface gravities works well for nearby stars for which distances are accurately known from Hipparcos parallaxes. For more distant late-type stars, the gravity can be determined spectroscopically from the difference in $[\text{Fe}/\text{H}]$ derived from neutral and ionized iron lines. Fe I lines change very little with g , whereas Fe II lines change significantly. Departures from LTE in the ionization equilibrium of Fe should, however, be taken into account. This may be done by requiring that the difference $[\text{Fe}/\text{H}](\text{Fe II}) - [\text{Fe}/\text{H}](\text{Fe I})$ has the same value as in the case of a standard star with a surface gravity that is accurately determined from (2.6). In this way, differential values of $\log g$ can be determined to a precision of about ± 0.05 dex (Nissen 2008).

Due to the non-LTE effects on the ionization balance of Fe, $[\text{Fe}/\text{H}]$ should be determined from Fe II lines because they represent the dominating ionization stage of iron. If LTE is assumed, $[\text{Fe}/\text{H}]$ derived from Fe I lines turns out to be 0.1–0.2 dex lower than $[\text{Fe}/\text{H}]$ derived from Fe II lines in the case of metal-poor F, G, and K stars. It is likely that this problem is due to a higher degree of Fe I ionization than predicted by Saha's equation (Mashonkina et al. 2011).

2.5 Diffusion and Dust-Gas Separation of Elements

As mentioned in Sect. 1, it is generally assumed that the atmosphere of a late-type star with an upper convection zone has retained a “fossil” record of the composition of the Galaxy at

the time and the place for the formation of the star. A high-resolution study of the metal-poor ($[\text{Fe}/\text{H}] \sim -2$) globular cluster NGC 6397 by Korn et al. (2007) indicates, however, that the abundances of Mg, Ca, Ti, and Fe in main-sequence turnoff stars are about 0.12 dex (30%) lower than the abundances of these elements in K giants. This may be explained by downward diffusion of the elements at the bottom of the convection zone for turnoff stars. The elements are depleted by about the same factor, so the effect of diffusion on abundance ratios is less than 0.05 dex.

For the solar atmosphere, the depletion by diffusion of elements heavier than boron is predicted to have been about 0.04 dex (Turcotte and Schweingruber 2002), and the effect of diffusion on abundance ratios is negligible. This is confirmed by the good agreement of abundance ratios for non-volatile elements in the solar atmosphere and in the most primitive meteorites, the carbonaceous chondrites (Asplund et al. 2009). A very precise study of solar “twin” stars (i.e., stars having nearly the same T_{eff} , g , and $[\text{Fe}/\text{H}]$ as the Sun) by Meléndez et al. (2009) shows, on the other hand, that the Sun has a higher abundance ratio of volatile elements (C, N, O, S, and Zn) with respect to Fe than the large majority of twin stars. The deviation is about 0.05 dex. As suggested by the authors, this may be explained by selective accretion of refractory elements, including iron, on dust particles in the protosolar disk. Thus, some fraction of the refractory elements may end up in terrestrial planets. If true, abundance ratios like $[\text{O}/\text{Fe}]$ and $[\text{S}/\text{Fe}]$ can deviate by ~ 0.05 dex from the original ratio in the interstellar cloud that formed the star depending on whether the star is with or without terrestrial planets.

It is concluded that the effects of diffusion may change some abundance ratios by up to 0.05 dex for metal-poor stars with relatively thin convection zones. In the case of disk stars, the abundances of volatile elements relative to iron could be affected by dust-gas separation in connection with star and planet formation by ~ 0.05 dex. For refractory elements, there are no indications of differences in abundance ratios between stars with and without detected planets (Neves et al. 2009).

3 Elements Used as Stellar Population Tracers

This section presents a discussion of some abundance ratios that have been used as tracers of stellar populations. In several cases, the nucleosynthesis and chemical evolution of the corresponding elements are not well understood; nevertheless, the abundance ratios have proven to be very useful in disentangling stellar populations. In addition, the observed differences and trends provide important constraints on SNe modeling and theories of Galactic chemical evolution.

3.1 Carbon and Oxygen

Abundances of C and O can be determined from spectral lines corresponding to forbidden transitions between low-excitation states and allowed transitions between high-excitation states for neutral atoms. In addition, molecular CH and OH lines in the blue-UV and infrared spectral regions can be applied.

The most reliable C and O abundances are derived from the forbidden $[\text{C I}] \lambda 8727$ and $[\text{O I}] \lambda 6300$ lines, provided that these rather weak lines are measured with sufficient spectral resolution and S/N . Both lines are blended, $[\text{O I}]$ by a Ni I line and $[\text{C I}]$ by a weak Fe I

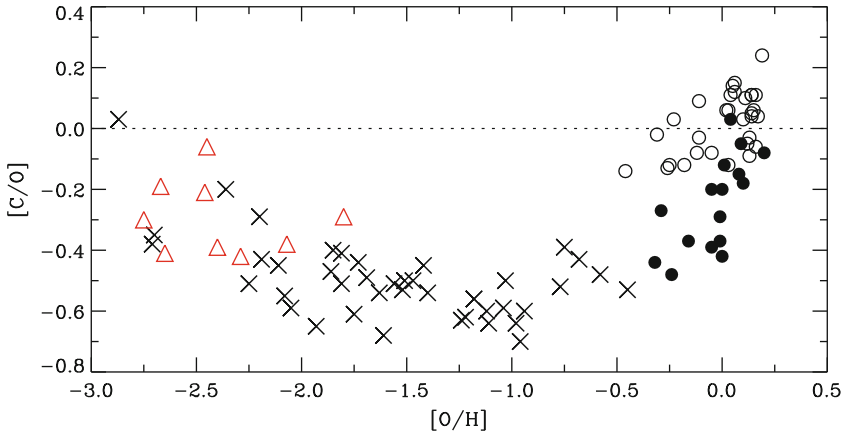
line. These blends must be taken into account when deriving abundances. A strong collisional coupling with the ground states ensures that the [C I] and [O I] lines are formed in LTE. The correction for 3D effects is small for the Sun (Asplund 2005) but increases with decreasing metallicity and may reach as much as -0.2 dex in turnoff stars with $[\text{Fe}/\text{H}] = -2$ (Nissen et al. 2002).

The [C I] line is too weak to be a useful abundance indicator for metal-poor ($[\text{Fe}/\text{H}] < -1$) dwarf and subgiant stars. In giants, carbon abundances are changed by dredge-up of gas affected by CN-cycle hydrogen burning. The [O I] line, on the other hand, can be used to derive oxygen abundances in dwarfs and subgiants down to $[\text{Fe}/\text{H}] \sim -2$ and in giants down to metallicities around -3 . Alternatively, carbon and oxygen abundances in halo stars can be derived from high-excitation atomic lines (the C I lines around $9,100 \text{ \AA}$ and the O I triplet at $7,774 \text{ \AA}$), but in both cases the non-LTE effects are uncertain (Fabbian et al. 2009). CH and OH lines can be used to derive carbon and oxygen abundances even at extremely low metallicities, but they are very sensitive to T_{eff} and large 3D corrections should be applied (Asplund 2005). Due to these problems, C and O abundances in halo stars are quite uncertain. Carbon seems to follow iron, i.e., $[\text{C}/\text{Fe}] \simeq 0$ from $[\text{Fe}/\text{H}] = 0$ to -3 (Bensby and Feltzing 2006; Fabbian et al. 2009). [O/Fe] raises from zero to about $+0.5$ dex, when the metallicity of disk stars decreases from $[\text{Fe}/\text{H}] = 0$ to -1 and then stays approximately constant at $[\text{O}/\text{Fe}] \simeq +0.5$ down to $[\text{Fe}/\text{H}] \simeq -2$ (Nissen et al. 2002). According to Cayrel et al. (2004), who derived oxygen abundances in giants from the [O I] line, the constant level of [O/Fe] continues all the way down to $[\text{Fe}/\text{H}] \simeq -3.5$, if one assumes that 3D corrections are the same as in metal-poor dwarf stars.

Although the trends of [C/Fe] and [O/Fe] as a function of [Fe/H] are somewhat uncertain due to non-LTE and 3D effects, the ratio between the abundances of C and O is more immune to these problems. This stems from the fact that the forbidden [C I] and [O I] lines have about the same dependence of temperature and pressure. Hence, the derived [C/O] is insensitive to 3D effects. The same is the case for [C/O] derived from high-excitation atomic lines. Furthermore, as shown by Fabbian et al. (2009), the non-LTE corrections of C and O abundances derived from C I and O I lines tend to cancel, so that the trend of [C/O] vs. [O/H] is fairly independent of the choice of the hydrogen collision parameter, S_{H} (see [Sect. 2.3](#)). [Figure 2-2](#) shows this trend for $S_{\text{H}} = 1$, with the abundances determined from forbidden lines for disk stars and from high-excitation atomic lines for the halo stars.

Carbon is synthesized in stellar interiors by the triple- α process, but it is unclear which objects are the main contributors to the chemical evolution of carbon in the Galaxy. Type II SNe, Wolf-Rayet stars, intermediate- and low-mass stars in the planetary nebula phase, and stars at the end of the giant phase have been suggested. Oxygen, on the other hand, seems to be produced exclusively by α -capture on C in short-lived massive stars and is dispersed to the interstellar medium by type II SNe. Hence, the C/O ratio has the potential of being a good tracer of stellar populations. The separation between thin- and thick-disk stars in [Fig. 2-2](#) shows that this is indeed the case.

The approximately constant $[\text{C}/\text{O}] \simeq -0.5$ for halo stars with $[\text{O}/\text{H}]$ between -2.0 and -0.5 probably corresponds to the C/O yield ratio for massive stars. The increase in [C/O] for the thick-disk stars may be due to metallicity dependent winds from Wolf-Rayet stars, whereas the delayed production of carbon by low- and intermediate-mass stars can explain the higher [C/O] in thin-disk stars. The upturn of [C/O] at the lowest values of [O/H], which is also found for distant damped Lyman- α (DLA) galaxies (Cooke et al. 2011), could be due to enhanced carbon production by massive first-generation stars with extremely high rotation velocities (Chiappini et al. 2006).



■ Fig. 2-2

The carbon-to-oxygen ratio as a function of the oxygen abundance. *Open circles* refer to thin-disk and *filled circles* to thick-disk stars with data adopted from Bensby and Feltzing (2006). *Crosses* are halo stars from Fabbian et al. (2009), and *(red) triangles* show DLA data from Cooke et al. (2011)

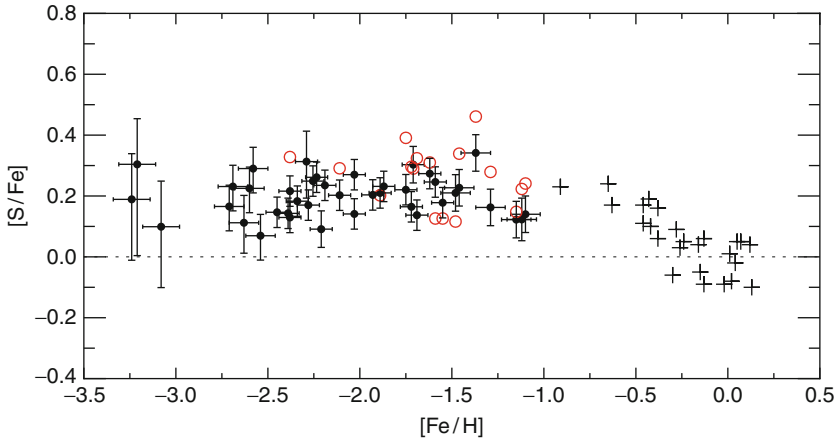
3.2 Intermediate-Mass Elements

The even- Z elements, Mg, Si, S, Ca, and Ti, are mainly produced by successive capture of α -particles in connection with carbon, oxygen, and neon burning in massive stars and dispersed into the interstellar medium by type II supernovae explosions on a timescale of $\sim 10^7$ years. Iron is also produced by SNe II, but the bulk of Fe comes from type Ia SNe on a much longer timescale ($\sim 10^9$ years). Hence, the ratio between the abundance of an α -capture element and iron in a star depends on how long the star formation process had proceeded before the star was formed. In this way, $[\alpha/\text{Fe}]$ becomes an important tracer of populations.

Mg, Si, Ca, and Ti abundances can be determined from several weak atomic lines in the optical spectra of late-type stars, whereas sulfur is more difficult, as discussed below. Traditionally, $[\alpha/\text{Fe}]$ is therefore defined as the average value of $[\text{Mg}/\text{Fe}]$, $[\text{Si}/\text{Fe}]$, $[\text{Ca}/\text{Fe}]$, and $[\text{Ti}/\text{Fe}]$ (see ● Sect. 2.1). As discussed by Asplund (2005), departures from LTE have some effects on the metallicity trends of $[\alpha/\text{Fe}]$ but probably not more than 0.1 dex, when weak lines are used. If the abundances are based on neutral lines, $[\alpha/\text{Fe}]$ is quite insensitive to T_{eff} and 3D effects, although high-excitation lines are to be preferred (Asplund 2005, Fig. 8).

The trend of $[\alpha/\text{Fe}]$ with metallicity is characterized by an increase of about 0.3 dex when $[\text{Fe}/\text{H}]$ decreases from 0 to -1 . Below $[\text{Fe}/\text{H}] = -1$, $[\alpha/\text{Fe}]$ is distributed around a plateau at 0.3 dex, but as discussed later, there are very significant differences in $[\alpha/\text{Fe}]$ at a given metallicity related to stellar populations both for disk and halo stars.

Sulfur is an α -capture element, and $[\text{S}/\text{Fe}]$ is therefore expected to show a plateau-like behavior for halo stars, but very high values $[\text{S}/\text{Fe}] \sim 0.8$ have been claimed at the lowest metallicities (Israeli and Rebolo 2001). These values may, however, be spurious due to the difficulty of measuring the very weak S I line at 8,694.6 Å. On the basis of stronger S I lines at 9,212.9 and 9,237.5 Å measured with UVES and corrected for telluric absorption lines, Nissen et al. (2007) find a plateau-like behavior of $[\text{S}/\text{Fe}]$, as shown in ● Fig. 2-3. Non-LTE corrections from Takeda et al. (2005) corresponding to $S_{\text{H}} = 1$ were included for sulfur, and the iron abundances were



■ Fig. 2-3

[S/Fe] as a function of [Fe/H]. *Plus signs* refer to data for disk stars from Chen et al. (2002) and *circles* to halo stars from Nissen et al. (2007). *Filled circles with error bars* are based on S abundances derived from the $\lambda\lambda 9212.9, 9237.5$ S I lines, whereas *open (red) circles* show data determined from the weak $\lambda 8694.6$ S I line

derived from Fe II lines with negligible non-LTE effects (Mashonkina et al. 2011). This result is supported by CRIRES observations of the near-IR S I triplet (see ● Fig. 2-1). Further studies of sulfur in Galactic stars would be important, especially because S is a volatile element that is undepleted onto dust. As such, it can be used to measure the α -enhancement of DLA galaxies.

In addition to the α -capture elements, Na is an interesting and sensitive tracer of stellar populations. It is thought to be made during carbon and neon burning in massive stars and is expelled by type II SNe together with the α -elements. The amount of Na made is, however, controlled by the neutron excess, which depends on the initial heavy element abundance in the star (Arnett 1971). This is probably the explanation of the fact that [Na/Mg] in halo stars correlates with [Mg/H] with a slope of about 0.5 (Nissen and Schuster 1997; Gehren et al. 2004).

Na abundances are best determined from the relatively weak Na I doublets $\lambda\lambda 5682.6, 5688.2$ and $\lambda\lambda 6154.2, 6160.7$, for which non-LTE and 3D corrections are rather small (Asplund 2005). At low metallicities, [Fe/H] < -2.0, these lines are too faint to be measured, and Na abundances are derived from the Na I D $\lambda\lambda 5890.0, 5895.9$ resonance lines. For this doublet, the non-LTE correction is large and reaches -0.4 dex for extremely metal-poor stars (Gehren et al. 2004). Furthermore, the use of these lines is sometimes complicated by overlapping interstellar Na I D lines and telluric H₂O lines.

In addition to Na production in massive stars, sodium can also be made in hydrogen-burning shells of intermediate- and low-mass stars via the CNO and Ne-Na cycles. This is probably the explanation of the Na-O anticorrelation in globular clusters, as discussed further in ● Sect. 6.3.

3.3 The Iron-Peak Elements

Among the iron-peak elements – Cr to Zn – the even-*Z* elements Cr, Fe, and Ni are represented by many lines in the spectra of late-type stars, which makes it possible to determine very precise abundance ratios, [Cr/Fe] and [Ni/Fe].

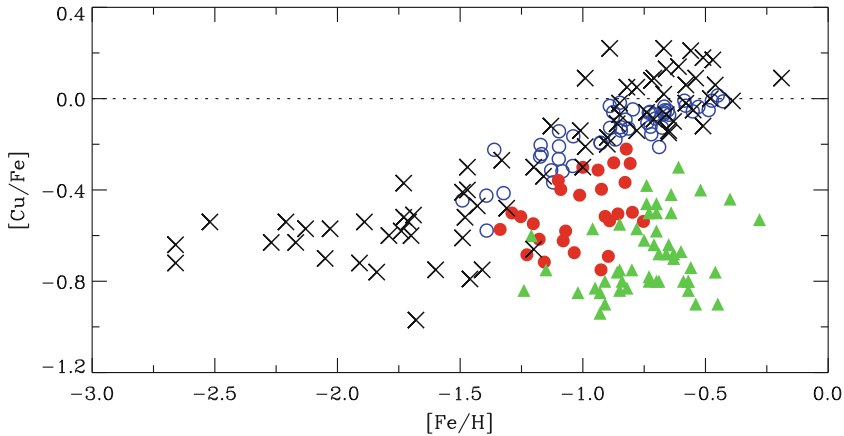
The ratio between Cr and Fe abundances is found to be the same as in the Sun, i.e., $[\text{Cr}/\text{Fe}] \simeq 0$, for all Galactic populations with $[\text{Fe}/\text{H}] > -2$. Below this metallicity, McWilliam et al. (1995) and Cayrel et al. (2004) have found a smooth decrease of $[\text{Cr}/\text{Fe}]$ to ~ -0.5 at $[\text{Fe}/\text{H}] = -4$ based on an LTE study of Cr I lines in very metal-poor giants. This is surprising, because Cr and Fe are predicted to be synthesized in constant ratios by explosive silicon burning in both types II and Ia SNe. The problem seems to have been solved by Bergemann and Cescutti (2010), who find that the derived decline of $[\text{Cr}/\text{Fe}]$ is an artifact caused by the neglect of non-LTE effects. They obtain a very satisfactory agreement between non-LTE Cr abundances derived from Cr I and Cr II lines when using surface gravities based on Hipparcos parallaxes.

For a long time, the Ni/Fe ratio was also thought to be solar in all stars. This is clearly the case for disk population stars (e.g., Chen et al. 2000), but Nissen and Schuster (1997, 2010) have shown that Ni is underabundant relative to Fe, i.e., $[\text{Ni}/\text{Fe}] \sim -0.1$ to -0.2 , in some halo stars with $[\text{Fe}/\text{H}] > -1.4$. These stars have also low $[\alpha/\text{Fe}]$ and $[\text{Na}/\text{Fe}]$ values, and a tight correlation between $[\text{Ni}/\text{Fe}]$ and $[\text{Na}/\text{Fe}]$ is present (see [Sect. 6.1](#)). Similar stars are found in dwarf galaxies (Tolstoy et al. 2009). The reason for this correlation is probably that the supernovae yields of ^{23}Na and ^{58}Ni depend on the neutron excess (see the detailed discussion by Venn et al. 2004).

The determination of abundances of the odd- Z elements Mn, Co, and Cu is more difficult than in the case of Cr, Fe, and Ni. There are fewer lines available and hyperfine-structure (HFS) splitting has to be taken into account. Furthermore, non-LTE effects seem to be very significant. Positive corrections of both $[\text{Mn}/\text{Fe}]$ (Bergemann and Gehren 2008) and $[\text{Co}/\text{Fe}]$ (Bergemann et al. 2010) are obtained, although the size of these corrections depends somewhat on the adopted S_{H} parameter for hydrogen collisions. This means that the strong decline of $[\text{Mn}/\text{Fe}]$ as a function of decreasing $[\text{Fe}/\text{H}]$ found from an LTE analysis of Mn I lines by, e.g., Reddy et al. (2006) and Neves et al. (2009) has to be modified to a more shallow trend. For Co, the non-LTE study of Bergemann et al. (2010) leads to surprisingly large over abundances of cobalt with respect to iron for halo stars, a result that is in disagreement with expectations from presently calculated supernovae yields.

Copper is a very interesting element as a tracer of metal-poor stellar populations. Abundances can be determined from the Cu I lines at 5,105.5, 5,218.2 and 5,782.1 Å. They are affected by HFS splitting to different degrees. Non-LTE calculations are not yet available, but LTE data show that $[\text{Cu}/\text{Fe}] \simeq 0$ for disk stars with no significant difference between the thin and the thick disk (Reddy et al. 2006). At $[\text{Fe}/\text{H}] \sim -1$, $[\text{Cu}/\text{Fe}]$ starts to decline steeply and reaches a plateau of $[\text{Cu}/\text{Fe}] \sim -0.6$ below $[\text{Fe}/\text{H}] \simeq -1.6$ (Mishenina et al. 2002) as shown in [Fig. 2-4](#). Low- α halo stars deviate, however, from this trend with a $[\text{Cu}/\text{Fe}]$ deficiency of 0.2–0.5 dex (Nissen and Schuster 2011). The same is the case for the more metal-rich part of stars in the globular cluster ω Centauri (Cunha et al. 2002) and stars belonging to the Sagittarius dSph galaxy (Sbordone et al. 2007) and the Large Magellanic Cloud (Pompéia et al. 2008), as shown in [Fig. 2-4](#). These $[\text{Cu}/\text{Fe}]$ data provide important constraints on the nucleosynthesis of Cu. Romano and Matteucci (2007) suggest that copper is initially made by explosive nucleosynthesis in type II SNe and later by a metallicity dependent neutron capture process (the weak s -process) in massive stars.

Zinc abundances can be determined from the $\lambda\lambda 4722.2, 4810.5$ Zn I lines. Non-LTE and 3D corrections are modest, i.e., of the order of +0.1 dex for metal-poor stars (Nissen et al. 2007). It is sometimes assumed that $[\text{Zn}/\text{Fe}] = 0$, which means that Zn can be used as a proxy of Fe in determining the metallicity of interstellar gas, e.g., in DLA systems; as a volatile element, Zn is not depleted onto dust like Fe. Newer studies show, however, that $[\text{Zn}/\text{Fe}]$ reaches +0.15 dex in metal-poor disk stars with perhaps a small separation between thin- and thick-disk stars (Bensby et al. 2005). For halo stars, $[\text{Zn}/\text{Fe}]$ increases from zero at $[\text{Fe}/\text{H}] = -1$ to $[\text{Zn}/\text{Fe}] \simeq 0.2$



■ Fig. 2-4

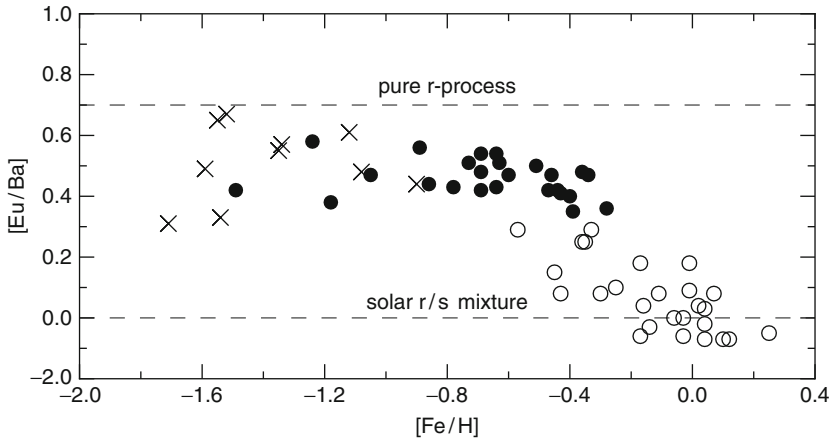
$[\text{Cu}/\text{Fe}]$ as a function of $[\text{Fe}/\text{H}]$. Stars studied by Mishenina et al. (2002) are indicated by crosses. Open (blue) circles refer to high- α stars and filled (red) circles to low- α halo stars from Nissen and Schuster (2011). Filled (green) triangles are data for giant stars in the inner part of the Large Magellanic Cloud adopted from Pompéia et al. (2008)

at $[\text{Fe}/\text{H}] = -2.5$ (Nissen et al. 2007), and at still lower metallicities, $[\text{Zn}/\text{Fe}]$ increases steeply to $[\text{Zn}/\text{Fe}] \simeq 0.5$ at $[\text{Fe}/\text{H}] = -3.5$ (Cayrel et al. 2004; Nissen et al. 2007). Hence, the behavior of zinc is complicated, and the nucleosynthesis of this element is not well understood. Several ways of producing Zn have been suggested: the weak s -process in massive stars, explosive Si burning in type II and type Ia SNe, and the main s -process in low- and intermediate-mass stars.

3.4 The Neutron Capture Elements

Among the heaviest elements, Y, Ba, and Eu are of particular interest as tracers of stellar populations. They are made by neutron capture processes, which are traditionally divided into the slow s -process and the rapid r -process. In the s -process, neutrons are added on a long timescale compared to that of β decays so that nuclides of the β -stability valley are built up. In the r -process, neutrons are added so fast that nuclides on the neutron-rich side of the stability valley are made. The s -process is divided into the main s -process that occurs in low- and intermediate-mass asymptotic giant branch (AGB) stars and the weak s -process occurring in massive stars. The r -process is not well understood, but it is thought to occur in connection with type II SNe explosions.

The relative contribution of the s - and the r -process to heavy elements in the solar system has been determined by Arlandini et al. (1999). Barium is called an s -process element because 81% of the solar Ba is due to the s -process. Europium, on the other hand, is an r -process element because 94% in the Sun originates from the r -process. In metal-poor stars, for which only massive stars have contributed to the nucleosynthesis of the elements, both Ba and Eu are, however, made by the r -process, provided that the contribution from the weak s -process is negligible. In such metal-poor stars, one would expect to find an r -process ratio between the abundances of europium and barium corresponding to $[\text{Eu}/\text{Ba}] \simeq 0.7$. These considerations suggest that



■ Fig. 2-5

[Eu/Ba] as a function of [Fe/H] with data adopted from Mashonkina et al. (2003). Crosses refer to stars with halo kinematics, filled circles to thick-disk stars, and open circles to thin-disk stars

[Eu/Ba] may be a useful tracer of stellar populations. As seen from Fig. 2-5, this is confirmed by Mashonkina et al. (2003).

Ba abundances can be determined from subordinate Ba II lines at 5,853.7, 6,141.7, and 6,496.9 Å. The stronger resonance line at 4,554.0 Å may also be used, but the analysis of this line is complicated by the presence of isotopic and HFS splitting. Europium abundances are primarily obtained from the Eu II line at 4,129.7 Å, which has to be analyzed by spectrum synthesis techniques, because the line is strongly broadened by isotopic and HFS splitting.

The barium-to-yttrium ratio is another interesting tracer of stellar populations. Yttrium ($Z = 39$) belongs to the first peak of s -process elements around the neutron magic number $N = 50$, whereas barium ($Z = 56$) is at the second peak around $N = 82$. The ratio Ba/Y (sometimes called heavy- s to light- s , hs/l s) depends on the neutron flux per seed nuclei and is predicted to be high for metal-poor, low-mass AGB stars (Busso et al. 1999). As pointed out by Venn et al. (2004), the Ba/Y ratio in stars belonging to dSph satellite galaxies is much higher than Ba/Y in Galactic halo stars with differences on the order of 0.6 dex in the metallicity range $-2 < [\text{Fe}/\text{H}] < -1$. Similar large offsets are found for stars with [Fe/H] around -1 in the globular cluster ω Cen (Smith et al. 2000) and for stars in the LMC (Pompéia et al. 2008). According to Fenner et al. (2006), this indicates that the chemical evolution of these systems has been so slow that winds from low-mass AGB stars have started to enrich the interstellar medium with s -process elements at a metallicity around [Fe/H] ~ -2 .

4 The Galactic Disk


4.1 The Thick and The Thin Disk


A long-standing problem in studies of Galactic structure and evolution has been the possible existence of a population of stars having kinematics, ages, and chemical abundances in between the characteristic values for the halo and the disk.

On the basis of a large program of $uvby-\beta$ photometry of F- and early G-type main-sequence stars within 100 pc, Strömberg (1987) concluded that an *intermediate Population II* does exist. $[\text{Fe}/\text{H}]$ was determined from a metallicity index $m_1 = (v - b) - (b - y)$, which is sensitive to the line blanketing in the v -band, and age was derived from the position of a star in the $c_1 - (b - y)$ diagram, where $c_1 = (u - v) - (v - b)$ is a measure of the Balmer discontinuity at 3,650 Å and hence of the surface gravity of the star. The color index $(b - y)$ is used as a measure of T_{eff} . After discussing the calibration of the m_1 index in terms of $[\text{Fe}/\text{H}]$, and c_1 as a function of absolute magnitude, Strömberg found that intermediate Population II consists of 10–15 Gyr old stars having $-0.8 < [\text{Fe}/\text{H}] < -0.4$ and velocity dispersions that are significantly greater than those of the younger, more metal-rich disk stars.

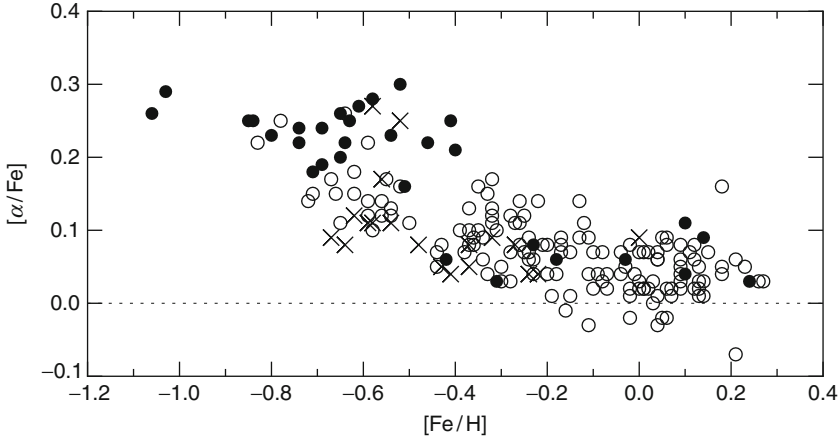
In a seminal paper, Gilmore and Reid (1983) showed that the distribution of stars in the direction of the Galactic South Pole cannot be fitted by a single exponential, but requires two disk components – a *thin disk* with a scale height of 300 pc and a *thick disk* with a scale height of about 1,300 pc. They identified intermediate Population II with the sum of the metal-poor end of the thin disk and the thick disk. Following this work, it has been intensively discussed if the thin and thick disks are discrete components of our Galaxy or if there is a more continuous sequence of stellar populations connecting the Galactic halo and the thin disk.

In another important paper, Edvardsson et al. (1993) derived precise abundance ratios for 189 stars belonging to the disk. Stars in the temperature range $5,600 < T_{\text{eff}} < 7,000$ K and somewhat evolved away from the zero-age main-sequence were selected from $uvby-\beta$ photometry of stars in the solar neighborhood (i.e., in a region around the Sun with a radius of ~ 100 pc) and divided into nine metallicity bins ranging from $[\text{Fe}/\text{H}] \sim -1.0$ to $\sim +0.3$. In each metallicity bin, the ~ 20 brightest stars were observed. Hence, there is no kinematical bias in the selection of the stars.

The Edvardsson et al. (1993) survey provides clear evidence for a scatter in $[\alpha/\text{Fe}]$ at a given metallicity for stars in the solar neighborhood. This is shown in  Fig. 2-6. As seen, $[\alpha/\text{Fe}]$ for stars in the metallicity range $-0.8 < [\text{Fe}/\text{H}] < -0.4$ is correlated with the mean galactocentric distance R_m in the stellar orbit. Stars with $R_m > 9$ kpc tend to have lower $[\alpha/\text{Fe}]$ than stars with $R_m < 7$ kpc, and stars belonging to the solar circle lie in between. Assuming that R_m is a statistical measure of the distance from the Galactic center at which the star was born, Edvardsson et al. explained the $[\alpha/\text{Fe}]$ variations as due to a star formation rate that declines with galactocentric distance. In other words, type Ia SNe start contributing iron at a higher $[\text{Fe}/\text{H}]$ in the inner parts of the Galaxy than in the outer parts.

The group of stars in Edvardsson et al. (1993) with $R_m < 7$ kpc have kinematical properties similar to those of the thick disk, for which the dispersions of the Galactic velocity components, U , V , and W , with respect to the local standard of rest (LSR) are determined to be $(\sigma_U, \sigma_V, \sigma_W) \simeq (65, 50, 40)$ km s $^{-1}$ and for which the asymmetric drift with respect to the LSR is $V_{\text{ad}} \simeq -50$ km s $^{-1}$. In comparison, thin-disk stars in the solar neighborhood have $(\sigma_U, \sigma_V, \sigma_W) \simeq (40, 30, 20)$ km s $^{-1}$ and $V_{\text{ad}} \simeq -10$ km s $^{-1}$. Thus, thick-disk stars move on more eccentric orbits than the thin-disk stars, and due to the increasing density of stars toward the inner part of the Galaxy, thick-disk stars presently situated in the solar neighborhood tend to be close to the apo-galactic distance in their orbits. This means that they have smaller mean galactocentric distances than the thin-disk stars. The differences in $[\alpha/\text{Fe}]$ shown in  Fig. 2-6 may, therefore, also be interpreted as due to a systematic difference in $[\alpha/\text{Fe}]$ between the thin and the thick disks. Apparently, Gratton et al. (1996) were first to suggest this interpretation of the $[\alpha/\text{Fe}]$ data.

A more clear chemical separation between thin- and thick-disk stars has been obtained by Fuhrmann (2004). For a sample of nearby stars with $5,300 < T_{\text{eff}} < 6,600$ K and



■ Fig. 2-6

$[\alpha/\text{Fe}]$ as a function of $[\text{Fe}/\text{H}]$ with data from Edvardsson et al. (1993). Stars shown with *filled circles* have a mean galactocentric distance in their orbits $R_m < 7$ kpc. *Open circles* refer to stars with $7 < R_m < 9$ kpc, and *crosses* refer to stars with $R_m > 9$ kpc

$3.7 < \log g < 4.6$, he derives very precise Mg abundances from Mg I lines and Fe abundances from Fe I and Fe II lines. In a $[\text{Mg}/\text{Fe}]$ vs. $[\text{Fe}/\text{H}]$ diagram, stars with thick-disk kinematics have $[\text{Mg}/\text{Fe}] \simeq +0.4$ and $[\text{Fe}/\text{H}]$ between -1.0 and -0.3 . The thin-disk stars show a well-defined sequence from $[\text{Fe}/\text{H}] \simeq -0.6$ to $+0.4$ with $[\text{Mg}/\text{Fe}]$ decreasing from $+0.2$ to 0.0 . Hence, there is a $[\text{Mg}/\text{Fe}]$ separation between thick- and thin-disk stars in the overlap region $-0.6 < [\text{Fe}/\text{H}] < -0.3$ with only a few “transition” stars. This is even more striking in a diagram, where $[\text{Fe}/\text{Mg}]$ is plotted as a function of $[\text{Mg}/\text{H}]$ (Fuhrmann 2004, Fig. 34).

On the basis of stellar ages derived from evolutionary tracks in the $M_{\text{bol}} - \log T_{\text{eff}}$ diagram, Fuhrmann (2004) finds that the maximum age of thin-disk stars is about 9 Gyr, whereas thick-disk stars have ages around 13 Gyr. This suggests that the systematic difference of $[\text{Mg}/\text{Fe}]$ is connected to a hiatus in star formation between the thick- and thin-disk phases.

4.2 The $[\alpha/\text{Fe}]$ Distribution of Disk Stars

Two major studies of abundance ratios in thin- and thick-disk stars (Bensby et al. 2005; Reddy et al. 2003, 2006) are based on kinematically selected groups of stars in the solar neighborhood. In these works, it is assumed that the kinematics of stars can be represented by Gaussian distribution functions for the velocity components, U , V , and W , with respect to the LSR. The kinematical probability that a star belongs to a given population: thin disk, thick disk or halo ($i = 1, 2, 3$), is then given by

$$P_i \propto k_i f_i \exp \left(-\frac{U^2}{2\sigma_{U_i}^2} - \frac{(V - V_{\text{adi}})^2}{2\sigma_{V_i}^2} - \frac{W^2}{2\sigma_{W_i}^2} \right), \quad (2.7)$$

where $k = (2\pi)^{-3/2} (\sigma_U \sigma_V \sigma_W)^{-1}$ is the standard normalization constant and f the relative number of stars in a given population. σ_U , σ_V , and σ_W are the velocity dispersions in U , V , and W , respectively, and V_{ad} the asymmetric drift velocity for the population. As an example, the

values used by Reddy et al. (2006) are given in [Table 2-1](#). Some of these values have considerable uncertainties. Depending on how the populations are defined, the fraction of thick-disk stars in the solar neighborhood may be as high as 20% (Fuhrmann 2004), and the local fraction of halo stars is often estimated to be on the order of 0.001.

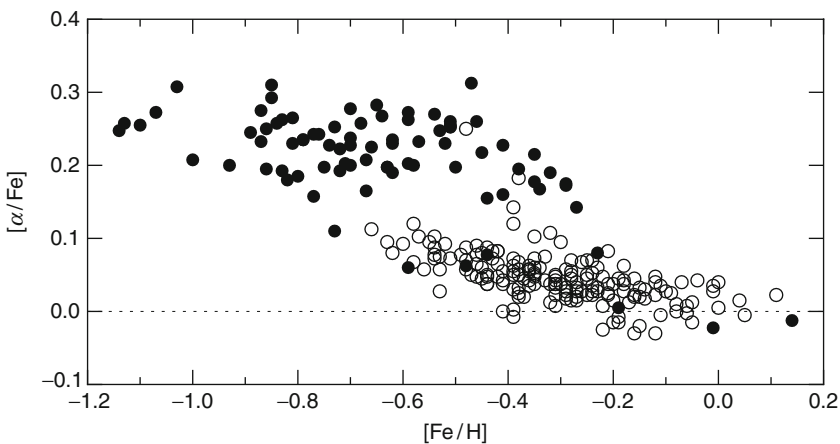
In the works of Bensby et al. (2005) and Reddy et al. (2003, 2006), precise abundance ratios for F and G dwarfs have been derived for samples of stars kinematically selected to have high probability of belonging to either the thin or the thick disk. Thus, in Reddy et al. (2006), the probability limit for each population is $P > 70\%$. The resulting $[\alpha/\text{Fe}]$ – $[\text{Fe}/\text{H}]$ diagram is shown in [Fig. 2-7](#). As seen, there is a gap in $[\alpha/\text{Fe}]$ between thin- and thick-disk stars for the metallicity range $-0.7 < [\text{Fe}/\text{H}] < -0.4$. Around $[\text{Fe}/\text{H}] = -0.3$, a few stars seem to connect the two populations like in the corresponding diagram of Fuhrmann (2004).

The $[\alpha/\text{Fe}]$ diagram of Bensby et al. (2005) looks similar to [Fig. 2-7](#) except that their thick-disk stars continue all the way up to solar metallicity with decreasing values of $[\alpha/\text{Fe}]$. Thus, Bensby et al. claim that star formation in the thick disk continued long enough to include chemical enrichment from type Ia SNe. This is, however, not so evident from the data of

■ Table 2-1

Velocity dispersions, asymmetric drift, and fraction of stars in the solar neighborhood for the thin-disk, thick-disk, and halo populations, as adopted by Reddy et al. (2006) for calculating membership probabilities

Population	σ_U km s ⁻¹	σ_V km s ⁻¹	σ_W km s ⁻¹	V_{ad} km s ⁻¹	f
Thin disk	43	28	17	-9	0.93
Thick disk	67	51	42	-48	0.07
Halo	131	106	85	-220	0.006



■ Fig. 2-7

$[\alpha/\text{Fe}]$ as a function of $[\text{Fe}/\text{H}]$ according to Reddy et al. (2003, 2006). Stars indicated by *open circles* have a probability $P > 70\%$ of belonging to the thin disk, whereas stars represented by *filled circles* have $P > 70\%$ of belonging to the thick disk

Reddy et al. (2006). In **Fig. 2-7**, the thick disk terminates around $[\text{Fe}/\text{H}] \simeq -0.3$ with little or no decrease of $[\alpha/\text{Fe}]$.

Given that the thin- and thick-disk stars of Bensby et al. (2005) and Reddy et al. (2003, 2006) have been kinematically selected, the question arises if there are stars with intermediate kinematics filling the gap in $[\alpha/\text{Fe}]$ between the two populations. The problem has been addressed by Ramírez et al. (2007), who derived oxygen abundances for 523 nearby stars from a non-LTE analysis of the O I triplet at 7,774 Å. A similar splitting as in **Fig. 2-7** between stars with thin- and thick-disk kinematics is obtained, and stars with intermediate kinematics do not fill the gap in $[\text{O}/\text{Fe}]$; they have either high $[\text{O}/\text{Fe}]$ or low $[\text{O}/\text{Fe}]$.

Another set of abundance data that can be used to study the problem of a gap in $[\alpha/\text{Fe}]$ between thin- and thick-disk stars has been obtained by Neves et al. (2009) from ESO/HARPS high-resolution spectra of 451 F, G, and K main-sequence stars in the solar neighborhood. The main purpose of this project is to detect planets around stars by measuring radial velocity variations with a precision of 1 m s^{-1} . As a by-product, stellar abundance ratios relative to those of the Sun have been derived.

Neves et al. (2009) show that trends of abundance ratios as a function of $[\text{Fe}/\text{H}]$ are the same for stars with and without planets. For both groups, there is a bifurcation of the abundances of α -capture elements relative to iron. This is shown in **Fig. 2-8**, where only stars having T_{eff} within $\pm 300 \text{ K}$ from the effective temperature of the Sun have been included in order to obtain a very high precision of $[\alpha/\text{Fe}]$, i.e., $\sigma [\alpha/\text{Fe}] \simeq \pm 0.02$. At metallicities $[\text{Fe}/\text{H}] < -0.1$, there is a gap in $[\alpha/\text{Fe}]$ between “high- α ” and “low- α ” stars. Hence, the data of Neves et al. (2009) confirm the dichotomy in $[\alpha/\text{Fe}]$ found by Bensby et al. (2005) and Reddy et al. (2006). Furthermore, the data of Neves et al. support the claim of Bensby et al. (2005) that the thick-disk stars have metallicities stretching up to solar metallicity.

The sample of stars from Neves et al. (2009) is *volume limited*; hence, the gap between high- α and low- α stars cannot be due to exclusion of stars with intermediate kinematics.

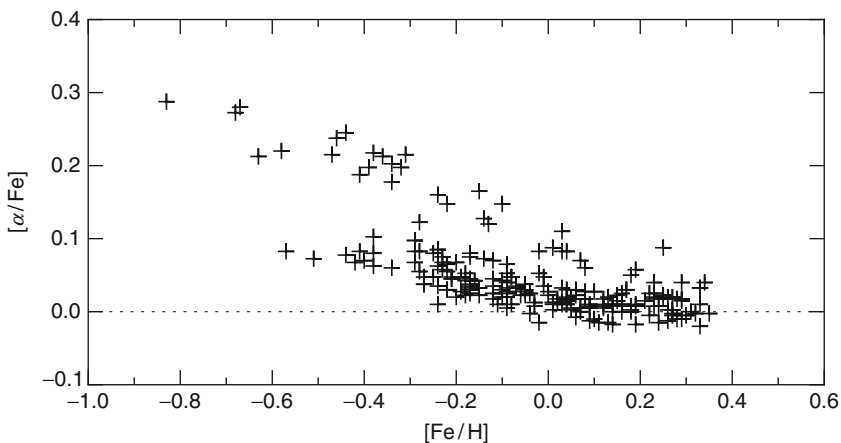
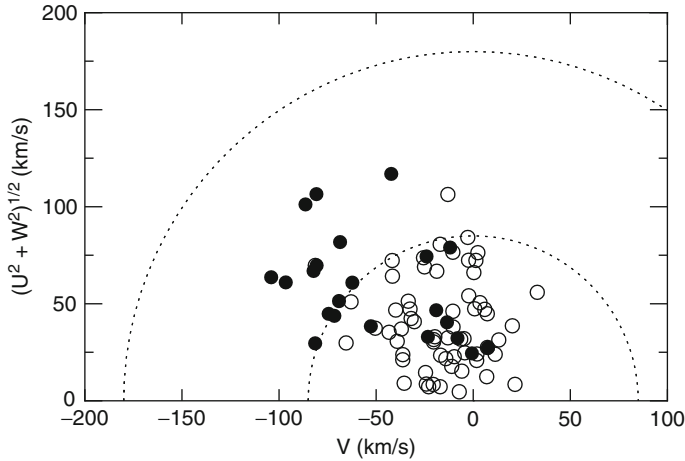


Fig. 2-8

The $[\alpha/\text{Fe}]$ vs. $[\text{Fe}/\text{H}]$ distribution for a *volume-limited* sample of main-sequence stars from Neves et al. (2009). Only stars with T_{eff} within $\pm 300 \text{ K}$ from the effective temperature of the Sun have been included



■ Fig. 2-9

The Toomre diagram for stars from [Fig. 2-8](#) with $[\text{Fe}/\text{H}] < -0.1$. High- α stars are shown with *filled circles* and low- α stars with *open circles*. The two circles delineate constant total space velocities with respect to the LSR, $V_{\text{tot}} = 85$ and 180 km s^{-1} , respectively

◆ [Figure 2-9](#) shows the distribution of the two populations in a Toomre energy diagram. As seen, the majority of high- α stars have a total space velocity with respect to the LSR larger than 85 km s^{-1} , which classify them as thick-disk stars, but several high- α stars are kinematically mixed with the low- α stars. This is an example of how an abundance ratio is a more clean separator of stellar populations than kinematics.

Considering that the high- α , thick-disk stars tend to be older than the oldest of the low- α , thin-disk stars (Fuhrmann 2004; Reddy et al. 2006), the difference in $[\alpha/\text{Fe}]$ is often explained in a scenario, where a period of rapid star formation in the early Galactic disk was interrupted by a merging satellite galaxy that “heated” the already formed stars to thick-disk kinematics. This was followed by a hiatus in star formation, in which metal-poor gas was accreted and type Ia SNe caused $[\alpha/\text{Fe}]$ to decrease. When star formation resumed, the first thin-disk stars were formed with low metallicity and low $[\alpha/\text{Fe}]$. This scenario also explains the systematic differences in $[\text{C}/\text{O}]$ (◆ [Fig. 2-2](#)) and $[\text{Eu}/\text{Ba}]$ (◆ [Fig. 2-5](#)) between thin- and thick-disk stars.

Haywood (2008) has pointed out that the low-metallicity thin-disk stars in the solar neighborhood tend to have positive values of the V velocity component. Such stars have mean galactocentric distances larger than the distance of the Sun from the Galactic center, and hence they are likely to have been formed in the outer part of the Galactic disk. Thick-disk stars, on the other hand, have negative values of V and tend to be formed in the inner Galactic disk. From these considerations, Haywood suggests that the bimodal distribution in $[\alpha/\text{Fe}]$ may be due to radial mixing of stars in the disk.


This scenario has been further investigated by Schönrich and Binney (2009a, b) in a model for the chemical evolution of the Galactic disk, for which the star formation rate is monotonically decreasing, and which includes radial migration of stars and gas flows. The model successfully fit the metallicity distribution and the large scatter in the age-metallicity relation for stars in the Geneva-Copenhagen survey (Nordström et al. 2004). The model also predicts a bimodal distribution of $[\alpha/\text{Fe}]$ for stars in the solar neighborhood with the high- α stars coming

from the inner parts of the Galactic disk and the low- α stars from the outer part. It will, however, be interesting to see if the model can reproduce a gap in the $[\alpha/\text{Fe}]$ distribution for disk stars, as found from the data of Neves et al. (2009).

4.3 Abundance Gradients in the Disk

In models for the chemical evolution of the Galactic disk, observed abundance gradients provide important constraints. Gradients may be determined from B-type stars and H II regions, but the most precise results have been obtained from Cepheids. These variable stars are bright enough to be studied at large distances, and accurate values of the distances can be obtained from the period-luminosity relation. Earlier results suggested a steeper metallicity gradient in the inner part of the Galaxy as compared to the outer part with a break in the gradient occurring around 10 kpc. According to newer work, this is, however, not so obvious. For 54 Cepheids ranging in galactocentric distance R_G from 4 to about 14 kpc, Luck et al. (2006) find an overall gradient $d[\text{Fe}/\text{H}]/dR_G = -0.06 \text{ dex kpc}^{-1}$, but there seems to be a significant cosmic scatter around a linear fit to the data. A region at Galactic longitude $l \sim 120^\circ$ and a distance of 3–4 kpc from the Sun has enhanced metallicities with $\Delta [\text{Fe}/\text{H}] \simeq +0.2$ in the mean. Such spatial inhomogeneities could be due to recent SNe events.

Yong et al. (2006) have made an interesting study of 24 Cepheids in the outer Galactic disk based on high-resolution spectra. The distance range is $12 < R_G < 18$ kpc. Most of the Cepheids continue the trend with galactocentric distance exhibited by the Luck et al. (2006) sample, but a minority of six Cepheids have $[\text{Fe}/\text{H}]$ around -0.8 dex and enhanced α -element abundances, $[\alpha/\text{Fe}] \sim +0.3$. Thus, there is some evidence for two populations of Cepheids in the outer disk.

Cepheids can only provide information about “present-day” gradients in the disk. Abundances of stars in open clusters may, however, be used to determine gradients at different ages. In this connection, one benefits from the quite accurate ages and distances that can be determined from color-magnitude diagrams of open clusters. For a review of results from clusters, the reader is referred to Friel ( Chap. 7, Open Clusters and their Role in the Galaxy).

5 The Galactic Bulge

Due to a large distance and a high degree of interstellar absorption and reddening, the Galactic bulge is the least known component of the Milky Way. It has been much discussed if the bulge contains only very old stars or if it also includes a younger population. This problem is related to two different scenarios for the formation of the bulge, a “classical” bulge formed rapidly by the coalescence of star-forming clumps, as suggested from the simulations of Elmegreen et al. (2008), or a “pseudobulge” formed over a longer time through dynamical instabilities in the Galactic disk (see review by Kormendy and Kennicutt 2004). The measurement of abundance ratios in bulge stars may help to decide between these scenarios by providing information on the timescale for the formation of the bulge.


The metallicity distribution of bulge stars has been debated for a long time. With the determination of $[\text{Fe}/\text{H}]$ for 800 bulge giants based on VLT multi-fiber spectra with a resolution of $R \sim 20,000$ (Zoccali et al. 2008), a robust result seems to have been obtained. Zoccali et al. observed stars in four fields having distances from the Galactic center ranging from 600 to 1,600 pc. The overall metallicity distribution is centered on solar metallicity and extends from


$[\text{Fe}/\text{H}] \simeq -1.5$ to $+0.5$, but with few stars in the range $-1.5 < [\text{Fe}/\text{H}] < -1.0$. A decrease of the mean metallicity along the bulge minor axis is suggested corresponding to a gradient of ~ 0.6 dex per kpc.

A pioneering study of α/Fe ratios in 12 bulge red giants was carried out by McWilliam and Rich (1994) based on 4-m telescope optical echelle spectra with a resolution of $R \sim 20,000$ and typical $S/N \sim 50$. Enhanced values of $[\text{Mg}/\text{Fe}]$ and $[\text{Ti}/\text{Fe}]$ ($\sim +0.3$ dex) were found for all stars up to a metallicity of $\sim +0.4$ dex, suggesting a very rapid formation of the bulge. In contrast, $[\text{Si}/\text{Fe}]$ and $[\text{Ca}/\text{Fe}]$ showed solar values, which is difficult to understand. These data are now superseded by higher-resolution spectra obtained with 8-m class telescopes both in the optical and the infrared spectral regions.

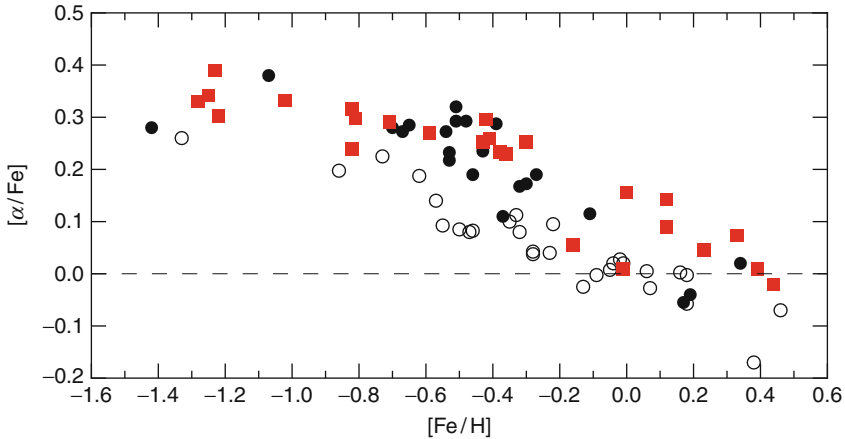
Lecureur et al. (2007) used VLT/UVES spectra in the spectral region 4,800–6,800 Å to derive O and Mg abundances relative to Fe for about 50 red bulge giants. They find that the $[\text{O}/\text{Fe}]$ and $[\text{Mg}/\text{Fe}]$ trends are more enhanced than the corresponding trends for thick-disk stars as determined by Bensby et al. (2005) for dwarf stars. The same conclusion is reached by Fulbright et al. (2007) from abundance ratios derived for 27 red giants observed at high resolution in the 5,000–8,000 Å spectral region with the Keck/HIRES spectrograph.

In contrast to these results, Meléndez et al. (2008) find the same $[\text{O}/\text{Fe}]$ – $[\text{Fe}/\text{H}]$ trend for 19 red giants belonging to the bulge and 21 thick-disk giants in the solar neighborhood. For both samples, the oxygen and iron abundances are based on spectrum synthesis of OH and Fe I lines in an infrared spectral region around 1.55 μm , as observed with Gemini/Phoenix. Hence, the comparison of the bulge and the thick disk is done in a differential way for the same type of stars. Meléndez et al. suggest that the comparisons of bulge giants with thick-disk dwarfs have led to a spurious offset of the two $[\text{O}/\text{Fe}]$ – $[\text{Fe}/\text{H}]$ trends due to systematic errors in the abundance ratios. When LTE is assumed and classical 1D model atmospheres are adopted, systematic errors could arise due to different non-LTE and 3D corrections for dwarf and giant stars. Recently, Ryde et al. (2010) have confirmed the results of Meléndez et al. (2008) from VLT/CRIRES high-resolution spectra of 11 red giants in the bulge.

Further evidence for an agreement between bulge and local thick-disk stars has been obtained by Alves-Brito et al. (2010) from optical high-resolution spectra. Their results are shown in  Fig. 2-10. As seen, the trends for bulge and thick-disk stars agree well with no significant differences below solar metallicity. The thin-disk stars, on the other hand, fall below the bulge and the thick-disk stars. Above solar metallicity, the bulge stars tend to have higher $[\alpha/\text{Fe}]$ values than the disk stars, but this needs to be confirmed for a larger sample.

Independent and very interesting data for abundances in the Galactic bulge have been obtained from spectra of microlensed main-sequence and subgiant stars. During a microlensing event, the flux from a star can be enhanced by a factor of 100 or even more; hence, the magnitude of a typical bulge turnoff star raises from $V \sim 18$ to $V \sim 13$. This makes it possible to obtain high-resolution spectra with good S/N . Bensby et al. (2011) have presented a homogeneous abundance study of 26 such microlensed stars using the same methods as applied for local thick-disk dwarf stars (Bensby et al. 2005). As seen from  Fig. 2-16, there is a tendency that these microlensed bulge stars are separated into two regions in the $[\alpha/\text{Fe}]$ – $[\text{Fe}/\text{H}]$ plane: A metal-poor group with $[\text{Fe}/\text{H}] < -0.3$ and $[\alpha/\text{Fe}] \simeq +0.3$ like the thick-disk stars and a metal-rich group with $+0.1 < [\text{Fe}/\text{H}] < +0.6$ for which the majority of stars have $[\alpha/\text{Fe}] \simeq 0$ like the thin-disk stars, but four stars have $[\alpha/\text{Fe}] \simeq +0.1$.

Another interesting aspect of the study of microlensed main-sequence and subgiant stars is the possibility to derive ages by comparing their position in a $\log g - \log T_{\text{eff}}$ diagram with isochrones computed from stellar models. According to Bensby et al. (2011), the metal-poor bulge stars have an average age of 11.2 Gyr with a dispersion of ± 2.9 Gyr much of which may be



■ Fig. 2-10

The $[\alpha/\text{Fe}]$ - $[\text{Fe}/\text{H}]$ relation for bulge K giant stars shown with filled (red) squares in comparison with K giants in the solar neighborhood having either thick-disk kinematics (filled circles) or thin-disk kinematics (open circles) (Data for $[\text{Mg}/\text{Fe}]$, $[\text{Si}/\text{Fe}]$, $[\text{Ca}/\text{Fe}]$, and $[\text{Ti}/\text{Fe}]$ are adopted from Alves-Brito et al. (2010))

due errors in the age determination. The metal-rich bulge stars are on average younger (7.6 Gyr) and has a larger age dispersion (± 3.9 Gyr).

On the basis of these new data, Bensby et al. (2011) suggest that the bulge consists of two stellar populations, i.e., a metal-poor population similar to the thick disk in terms of metallicity range, $[\alpha/\text{Fe}]$, and age and a metal-rich population, which could be related to the inner thin disk. Supporting evidence has recently been obtained by Hill et al. (2011) from a study of 219 red giants in Baade's bulge window (situated about 4° from the Galactic center). $[\text{Fe}/\text{H}]$ and $[\text{Mg}/\text{Fe}]$ are derived from ESO VLT FLAMES spectra with a resolution of $R = 20,000$. The distribution of $[\text{Fe}/\text{H}]$ is asymmetric and can be deconvolved into two Gaussian components: a metal-poor population centered at $[\text{Fe}/\text{H}] = -0.30$ having a dispersion of 0.25 dex in $[\text{Fe}/\text{H}]$ and a metal-rich population centered at $[\text{Fe}/\text{H}] = +0.32$ with a dispersion of 0.11 dex only. The metal-poor population has $[\text{Mg}/\text{Fe}] \simeq 0.3$, whereas the stars in the metal-rich population distribute around the solar Mg/Fe ratio. These data agree well with those of Bensby et al. (2011). However, a larger set of data for microlensed bulge stars and information about chemical abundances of inner disk stars are needed before one can draw any robust conclusions about the existence of two bulge populations and the consequences this may have for models of the formation of the bulge.

6 The Galactic Halo

6.1 Evidence of Two Distinct Halo Populations

For a long time, it has been discussed if the Galactic halo consists of more than one population. The classical monolithic collapse model of Eggen et al. (1962) corresponds to a single halo population, but from a study of globular clusters, Searle and Zinn (1978) suggested that the halo comprises two populations: (1) an inner, old, flattened population with a slight prograde

rotation formed during a dissipative collapse and (2) an outer, younger, spherical population accreted from satellite systems. This dichotomy of the Galactic halo has found support in a study of $\sim 20,000$ stars from the Sloan Digital Sky Survey (SDSS) by Carollo et al. (2007). They find that the inner halo consists of stars with a peak metallicity at $[\text{Fe}/\text{H}] \simeq -1.6$, whereas the outer halo stars distribute around $[\text{Fe}/\text{H}] \simeq -2.2$.

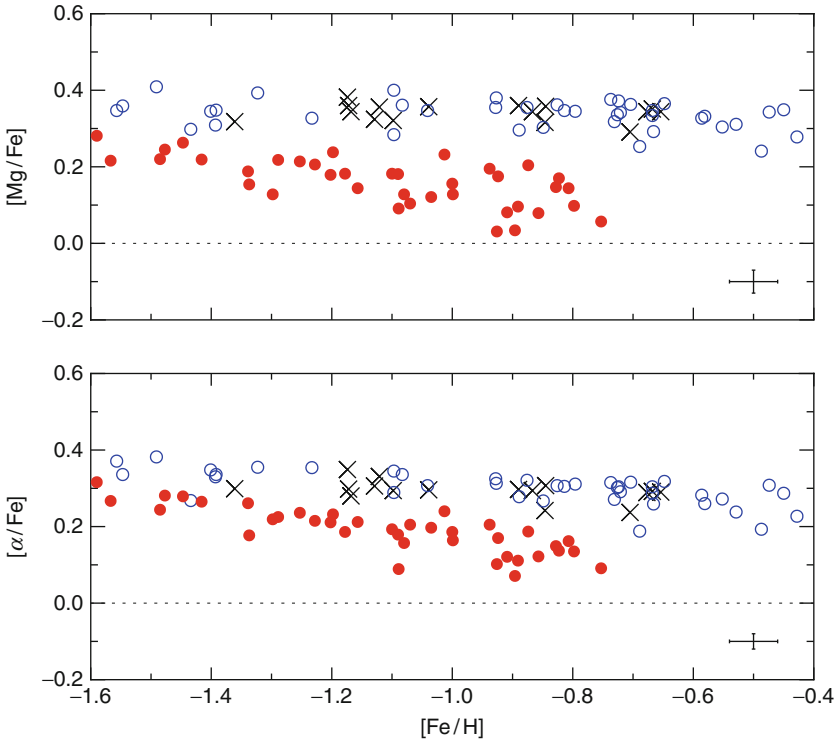
Several studies suggest that there is a difference in $[\alpha/\text{Fe}]$ between stars that can be associated with the inner and the outer halo, respectively. Fulbright (2002) shows that stars with large values of the total space velocity relative to the LSR, $V_{\text{tot}} > 300 \text{ km s}^{-1}$, tend to have lower values of $[\alpha/\text{Fe}]$ than stars with $150 < V_{\text{tot}} < 300 \text{ km s}^{-1}$, and Stephens and Boesgaard (2002) find that $[\alpha/\text{Fe}]$ is correlated with the apo-galactic orbital distance in the sense that the outermost stars have the lowest values of $[\alpha/\text{Fe}]$. Further support for such differences in $[\alpha/\text{Fe}]$ comes from a study by Gratton et al. (2003), who divided stars in the solar neighborhood into two populations according to their kinematics: (1) a “dissipative” component, which includes thick-disk stars and prograde rotating halo stars, and (2) an “accretion” component that consists of retrograde rotating halo stars. The “accretion” component has smaller values and a larger scatter for $[\alpha/\text{Fe}]$ than the “dissipative” component.

The differences in $[\alpha/\text{Fe}]$ found in these works are not larger than about 0.1 dex, and it is unclear if the distribution of $[\alpha/\text{Fe}]$ is continuous or bimodal. Nissen and Schuster (1997) found a more clear dichotomy in $[\alpha/\text{Fe}]$ for 13 halo stars with $-1.3 < [\text{Fe}/\text{H}] < -0.5$. Eight of these halo stars have $[\alpha/\text{Fe}]$ in the range 0.1–0.2 dex, whereas $[\alpha/\text{Fe}] \simeq 0.3$ for the other five halo stars. Interestingly, the low- α halo stars tend to have larger apo-galactic distances than the high- α stars.

A more extensive study of “metal-rich” halo stars has been carried out by Nissen and Schuster (2010). Stars are selected from the Schuster et al. (2006) $uvby-\beta$ catalogue of high-velocity and metal-poor stars. To ensure that a star has a high probability of belonging to the halo population, the total space velocity with respect to the LSR is required to be larger than 180 km s^{-1} . Furthermore, Strömgren photometric indices are used to select dwarfs and subgiants with $5,200 < T_{\text{eff}} < 6,300 \text{ K}$ and $[\text{Fe}/\text{H}] > -1.6$. High-resolution, high S/N spectra were obtained with the ESO VLT/UVES and the Nordic Optical Telescope FIES spectrographs for 78 of these stars. The large majority of the stars are brighter than $V = 11.1$ and situated within a distance of 250 pc. These spectra are used to derive high-precision LTE abundance ratios in a differential analysis that also includes 16 stars with thick-disk kinematics. The precision of the various abundance ratios ranges from 0.02 to 0.04 dex.

Figure 2-11 (top) shows the distribution of $[\text{Mg}/\text{Fe}]$ as a function of $[\text{Fe}/\text{H}]$ for stars in Nissen and Schuster (2010). As seen, the halo stars split into two distinct populations: “high- α ” stars with a nearly constant $[\text{Mg}/\text{Fe}]$ and “low- α ” stars with declining values of $[\text{Mg}/\text{Fe}]$ as a function of increasing metallicity. A classification into these two populations has been done on the basis of $[\text{Mg}/\text{Fe}]$, but as seen from the bottom part of the figure, $[\alpha/\text{Fe}]$ would have led to the same division of the halo stars except at the lowest metallicities, $-1.6 < [\text{Fe}/\text{H}] < -1.4$, where the two populations tend to merge, and the classification is less clear.

As seen from Figure 2-11, the separation in $[\text{Mg}/\text{Fe}]$ for the two halo populations is significantly larger than the separation in $[\alpha/\text{Fe}]$. At $[\text{Fe}/\text{H}] \simeq -0.8$, the mean difference in $[\text{Mg}/\text{Fe}]$ is about 0.25 dex, whereas it is only about 0.15 dex in $[\alpha/\text{Fe}]$. This is probably caused by different degrees of SNe Ia contribution to the production of Mg, Si, Ca, and Ti. According to Tsujimoto et al. (1995, Table 3), the relative contribution of SNe Ia to the solar composition is negligible for Mg, 17% for Si, and 25% for Ca (Ti was not included). For comparison, the SNe Ia contribution is 57% for Fe. Hence, $[\text{Mg}/\text{Fe}]$ is a more sensitive measure of the ratio between type II and



■ Fig. 2-11

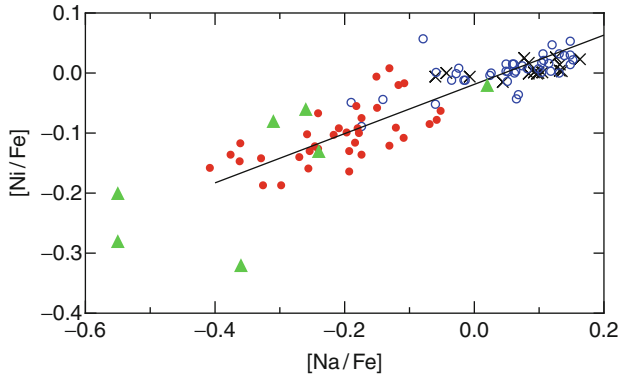
[Mg/Fe] and $[\alpha/\text{Fe}]$ vs. $[\text{Fe}/\text{H}]$ based on data from Nissen and Schuster (2010). Stars with thick-disk kinematics are indicated by crosses and halo stars by circles. On the basis of the $[\text{Mg}/\text{Fe}]$ distribution, the halo stars are divided into low- α stars shown by filled (red) circles, and high- α stars shown by open (blue) circles

type Ia contributions to the chemical enrichment of matter than $[\alpha/\text{Fe}]$. On the other hand, it is possible to measure $[\alpha/\text{Fe}]$ with a higher precision than $[\text{Mg}/\text{Fe}]$ because many more spectral lines can be used to determine $[\alpha/\text{Fe}]$.

The low- α halo stars also have low values of $[\text{Na}/\text{Fe}]$ and $[\text{Ni}/\text{Fe}]$ relative to the high- α stars, and as shown in [Fig. 2-12](#), the two populations are well separated in a $[\text{Ni}/\text{Fe}]$ - $[\text{Na}/\text{Fe}]$ diagram. As seen, some stars in dSph galaxies are even more extreme in $[\text{Na}/\text{Fe}]$ and $[\text{Ni}/\text{Fe}]$ than the low- α halo stars.

6.2 Kinematics and Origin of the Two Halo Populations

The distribution of $[\alpha/\text{Fe}]$ in [Fig. 2-11](#) can be explained if the high- α stars have been formed in regions with such a high rate of chemical evolution that only type II SNe have contributed to the chemical enrichment up to $[\text{Fe}/\text{H}] \sim -0.4$. The low- α stars, on the other hand, originate from regions with a relatively slow chemical evolution so that type Ia SNe have started to contribute iron around $[\text{Fe}/\text{H}] = -1.6$ causing $[\alpha/\text{Fe}]$ to decrease toward higher metallicities until $[\text{Fe}/\text{H}] \sim -0.8$.



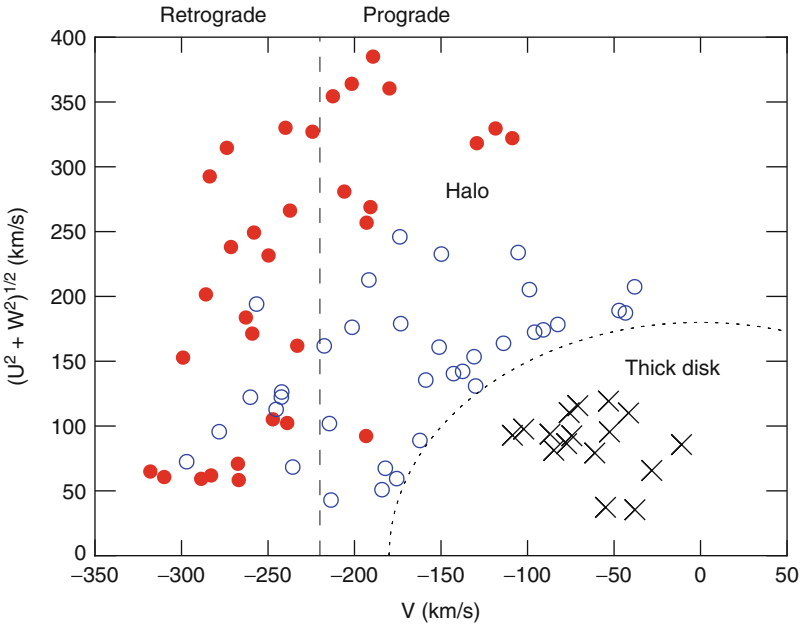
■ Fig. 2-12

The relation between $[\text{Ni}/\text{Fe}]$ and $[\text{Na}/\text{Fe}]$. Local main-sequence stars from Nissen and Schuster (2010) are indicated by the same symbols as in Fig. 2-11, whereas the (green) triangles refer to K giants in dSph satellite galaxies from Venn et al. (2004). For both sets of data, the stars are confined to the metallicity range $-1.6 < [\text{Fe}/\text{H}] < -0.4$

Further insight into the origin of the two halo populations can be obtained from kinematics. As seen from the Toomre energy diagram in Fig. 2-13, the high- α stars show evidence for being more bound to the Galaxy and favoring prograde Galactic orbits, while the low- α stars are less bound with two-thirds of them being on retrograde orbits. This suggests that the high- α population is connected to a dissipative component of the Galaxy, while the low- α stars have been accreted from dwarf galaxies.

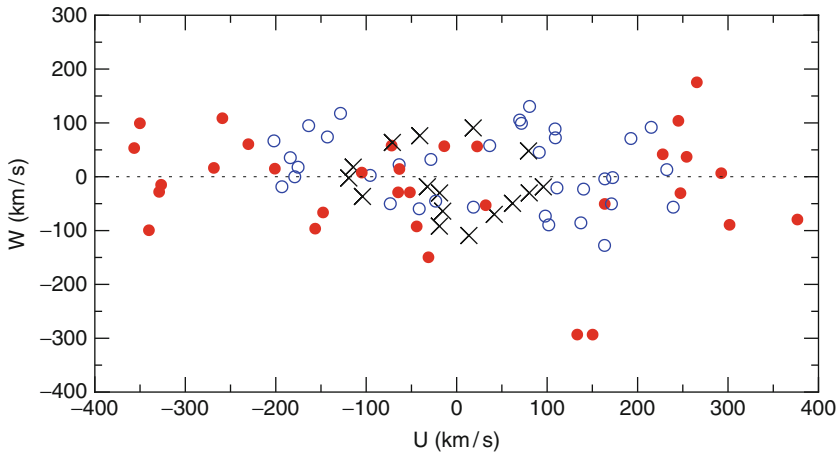
Several retrograde moving low- α stars have a Galactic V -velocity component similar to that of the ω Cen globular cluster, i.e., $V \sim -260 \text{ km s}^{-1}$. As often discussed (e.g., Bekki and Freeman 2003), ω Cen is probably the nucleus of a captured satellite galaxy with its own chemical enrichment history. Meza et al. (2005) have simulated the orbital characteristics of the tidal debris of such a satellite dragged into the Galactic plane by dynamical friction. The captured stars are predicted to have rather small W -velocities but a wide, double-peaked U distribution. As shown in Fig. 2-14, the W - U distribution observed for the low- α halo corresponds quite well to that prediction. There are two groups of low- α stars with $U > 200 \text{ km s}^{-1}$ and $U < -200 \text{ km s}^{-1}$, respectively, corresponding to stars moving in and out of the solar neighborhood on elongated radial orbits. Thus, a good fraction of the low- α stars, although not all, may well have originated in the ω Cen progenitor galaxy. The high- α stars, on the other hand, are confined to a much smaller range in U , i.e., from about -200 to about $+200 \text{ km s}^{-1}$.

The $[\alpha/\text{Fe}]$ vs. $[\text{Fe}/\text{H}]$ trend for the low- α stars in Fig. 2-11 and the Ni-Na trend in Fig. 2-12 resemble the corresponding trends for stars in dwarf galaxies. Stars in these systems tend, however, to have lower values of $[\alpha/\text{Fe}]$, $[\text{Na}/\text{Fe}]$, and $[\text{Ni}/\text{Fe}]$ than low- α halo stars. This offset agrees with simulations of the chemical evolution of a hierarchically formed stellar halo in a Λ CDM Universe by Font et al. (2006, Fig. 9). The bulk of halo stars originate from early accreted, massive dwarf galaxies with efficient star formation, whereas surviving satellite galaxies in the outer halo on average have smaller masses and a slower chemical evolution with a larger contribution from type Ia SNe at a given metallicity. The $[\text{Mg}/\text{Fe}]$ vs. $[\text{Fe}/\text{H}]$ trend for



■ Fig. 2-13

Toomre diagram for stars from Nissen and Schuster (2010) having $[\text{Fe}/\text{H}] > -1.4$. The same symbols for high- α halo, low- α halo, and thick-disk stars as in [Fig. 2-11](#) are used. The *short-dashed circle* corresponds to $V_{\text{tot}} = 180 \text{ km s}^{-1}$. The *long-dashed line* indicates zero rotation in the Galaxy and therefore separates retrograde moving stars from prograde moving



■ Fig. 2-14

The relation between the U and W velocity components for stars from Nissen and Schuster (2010) having $[\text{Fe}/\text{H}] > -1.4$

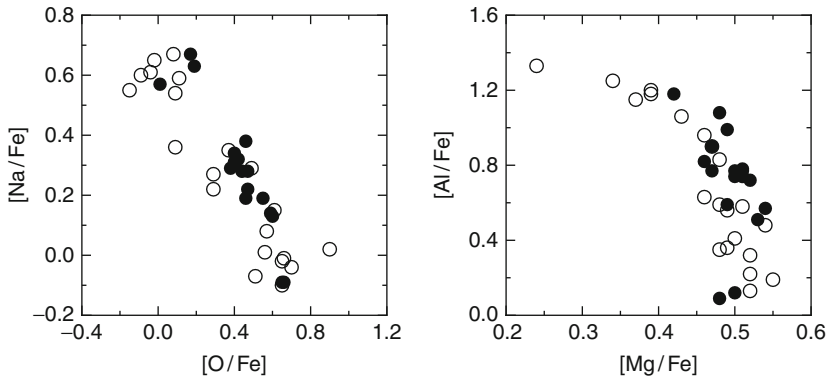
field stars, predicted by Font et al., agrees in fact remarkably well with the trend for the low- α halo stars.

The simulations of Font et al. (2006) do not explain the existence of high- α halo stars. Two recent Λ CDM simulations suggest, however, a dual origin of stars in the inner Galactic halo. Purcell et al. (2010) propose that ancient stars formed in the Galactic disk can be ejected to the halo by merging satellite galaxies, and Zolotov et al. (2009, 2010) find that stars formed out of accreted gas in the inner 1 kpc of the Galaxy can be displaced into the halo through a succession of mergers. Alternatively, the high- α population may simply belong to the high-velocity tail of a thick disk with a non-Gaussian velocity distribution.

6.3 Globular Clusters and Dwarf Galaxies

The Galactic halo contains more than 100 globular clusters and a number of dwarf spheroidal galaxies that are considered as Milky Way satellite systems. A dSph galaxy has a broad range in age and metallicity, whereas a globular cluster is a smaller system with a more limited range in these parameters. As often discussed, it is possible that some or all field halo stars come from dissolved globular clusters or have been accreted from previous generations of satellite galaxies. In this connection, it is of great interest to compare the chemical abundance ratios in field stars with the corresponding ratios in still existing globular clusters and dwarf galaxies.

Globular clusters were for a long time considered as examples of systems containing a single population of stars with a well-defined age and chemical composition. During the last decades, it has, however, become more and more clear that many, if not all, globular clusters contain multiple stellar populations. Strong evidence comes from abundance ratios between elements from oxygen to aluminum. As reviewed by Gratton et al. (2004), high-resolution spectroscopy of red giants has revealed anticorrelations between $[\text{Na}/\text{Fe}]$ and $[\text{O}/\text{Fe}]$ and between $[\text{Al}/\text{Fe}]$ and $[\text{Mg}/\text{Fe}]$ in intermediate metallicity globular clusters. An example is shown in [Fig. 2-15](#) for K giants in NGC 6752 (Yong et al. 2005)



■ Fig. 2-15

The Na-O and Al-Mg anticorrelations in the globular cluster NGC 6752 based on data from Yong et al. (2005). Filled circles indicate stars near the bright end of the red-giant branch; open circles refer to less luminous stars around the red-giant bump

The extensive study of Carretta et al. (2009) based on high-resolution VLT/UVES spectra for giant stars in 19 globular clusters indicates the existence of a $[\text{Na}/\text{Fe}]-[\text{O}/\text{Fe}]$ anticorrelation in all cases, but the amplitude of the variations is different from cluster to cluster. Variations in $[\text{Al}/\text{Fe}]$ correlated with $[\text{Na}/\text{Fe}]$ are seen in the majority of clusters, and an anticorrelation between $[\text{Al}/\text{Fe}]$ and $[\text{Mg}/\text{Fe}]$ is detected in a few cases. In addition, Yong et al. (2003) have revealed a correlation between the heavy magnesium isotope ^{26}Mg and Al in NGC 6752 by determining Mg isotope ratios from the profiles of MgH lines near 5,140 Å.

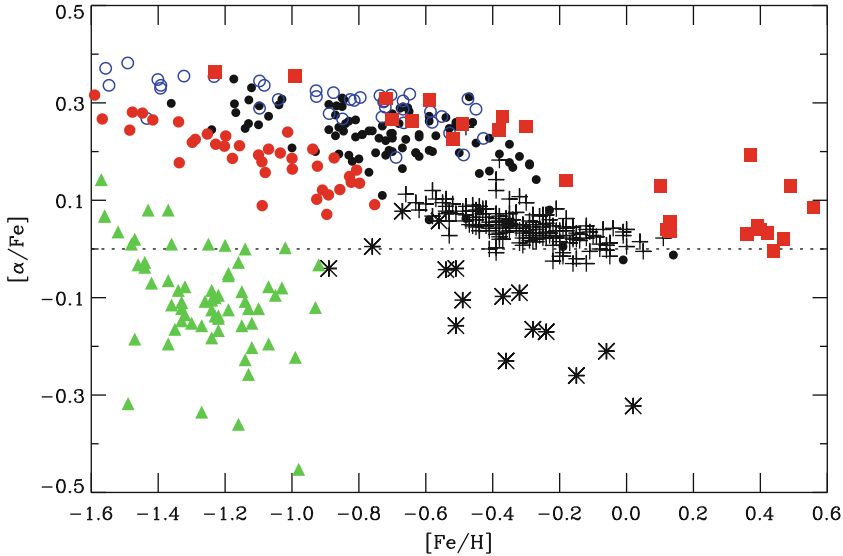
All of these abundance variations are ascribed to hydrogen burning via the CNO-cycle and the NeNa and MgAl chains. Some years ago, it was much discussed if the variations were due to nuclear processes and nonstandard mixing in the stars themselves or to an early generation of stars that have polluted the gas out of which the present low-mass stars in the clusters have formed. The fact that the abundance variations are not correlated with the luminosity of the giant stars speaks against the first possibility, and when Gratton et al. (2001) discovered that a Na-O anticorrelation is present in turnoff and subgiant stars in NGC 6397 and NGC 6752, the internal mixing case was ruled out. Hence, the abundance variations must be due to an early generation of stars. Candidates are intermediate-mass AGB stars undergoing hot-bottom burning (Ventura et al. 2001) and massive rotating stars (Decressin et al. 2007).

In addition to the abundance variations of elements from O to Al, there is also increasing evidence for variations of Ca, Fe, and *s*-process elements in several globular clusters. The classical example is ω Cen for which giant stars are found to have metallicities from $[\text{Fe}/\text{H}] \simeq -1.9$ to $\simeq -0.5$ (see review by Gratton et al. 2004). Multiple sequences in color-magnitude diagrams are present, suggesting the existence of four or five discrete populations in ω Cen with significant age and abundance differences. The more metal-rich stars in ω Cen have unusually low values of $[\text{Cu}/\text{Fe}]$ (Cunha et al. 2002) and very high values of the abundances of the second-peak *s*-process elements, i.e., $[\text{Ba}/\text{Fe}]$ and $[\text{La}/\text{Fe}]$ values around 1.0 dex (Smith et al. 2000). These abundance anomalies point to a complicated chemical evolution history, in which ω Cen was originally the nucleus of a much larger dwarf galaxy that merged with the early Galactic disk (Bekki and Freeman 2003).

Evidence for abundance variations of Ca, Fe, and *s*-process elements has been found for other clusters than ω Cen. On the basis of high-resolution spectroscopy with UVES of eight giant stars in NGC 1851, Yong and Grundahl (2008) found a range of 0.6 dex in $[\text{Zr}/\text{Fe}]$ and $[\text{La}/\text{Fe}]$, and Marino et al. (2009) found similar large variations in $[\text{Y}/\text{Fe}]$, $[\text{Zr}/\text{Fe}]$, and $[\text{Ba}/\text{Fe}]$ for 17 giant stars in M 22. In addition, there is a correlation between these ratios and $[\text{Fe}/\text{H}]$, which varies from -1.9 to -1.5 in M 22. Evidence of variations in $[\text{Ca}/\text{Fe}]$ is also found in other globular clusters by Lee et al. (2009) from photometric measurements of an index of the Ca II H and K lines, but Carretta et al. (2010) do not confirm this on the basis of high-resolution UVES spectra. Instead, they suggest that the spread in the Ca II H and K index may be related to variations in helium and nitrogen abundances.

Some authors, e.g., Lee et al. (2009), suggest that all globular clusters were once nuclei of dwarf galaxies that are now accreted and dissolved in the Milky Way. In this connection, one may wonder why practically no field halo stars share the Na-O abundance anomalies of the globular clusters. Perhaps the explanation is that the elements produced by AGB stars are confined to the potential wells of the clusters. According to the hydrodynamical simulations by D'Ercole et al. (2008), the gas ejected by AGB stars collects in cooling flows into the cores of globular clusters.

The chemical composition of giant stars in dSph galaxies is reviewed by Tolstoy et al. (2009). For the large majority of stars with $[\text{Fe}/\text{H}] > -2$, $[\alpha/\text{Fe}]$ is significantly lower than in halo and



■ Fig. 2-16

$[\alpha/\text{Fe}]$ vs. $[\text{Fe}/\text{H}]$ for various stellar populations. Thin-disk stars from Reddy et al. (2003) are shown with *plus symbols*. Filled circles refer to thick-disk stars from Reddy et al. (2006) and Nissen and Schuster (2010). Filled (red) squares are microlensed bulge stars from Bensby et al. (2011). Open (blue) circles are high- α and filled (red) circles low- α halo stars from Nissen and Schuster (2010). Asterisks refer to stars in the Sagittarius dSph galaxy (Sbordone et al. 2007), and filled (green) triangles show data for stars in the Sculptor dSph galaxy from Kirby et al. (2009) for which the precision of $[\alpha/\text{Fe}]$ is better than 0.15 dex

disk stars belonging to the Milky Way. Two examples are shown in ► Fig. 2-16. Stars in Sculptor (Kirby et al. 2009) distribute around $[\alpha/\text{Fe}] = -0.1$ and do not seem to have metallicities above $[\text{Fe}/\text{H}] \sim -0.9$. The other example is the Sagittarius dwarf galaxy, which was discovered by Ibata et al. (1994) to be merging with the Milky Way. The abundances derived by Sbordone et al. (2007) for giant stars show a declining trend of $[\alpha/\text{Fe}]$ reaching $[\alpha/\text{Fe}]$ as low as -0.3 dex at solar metallicity.

Evidently, present-day dSph galaxies have had a chemical evolution history different from that of the other main components of the Milky Way. In addition to the underabundance of $[\alpha/\text{Fe}]$, the more metal-rich stars in dSph galaxies stand out by having low $[\text{Na}/\text{Fe}]$ and $[\text{Ni}/\text{Fe}]$ abundances (► Fig. 2-11) as well as high $[\text{Ba}/\text{Y}]$ ratios (Venn et al. 2004). At the lowest metallicities, $[\text{Fe}/\text{H}] < -2$, the abundance ratios are, however, similar to those in Galactic halo stars, which suggests that the abundance deviations of dSph stars should not be explained in terms of an anomalous IMF. The most obvious reason for the underabundances of $[\alpha/\text{Fe}]$ is that the star formation rate in surviving dSph galaxies has been so slow that type Ia SNe started to contribute Fe at a metallicity of $[\text{Fe}/\text{H}] \sim -2$. Dwarf galaxies are, however, not ruled out as “building blocks” of the Galactic halo, because early accreted dSph galaxies probably have had a somewhat faster chemical evolution than the less massive present-day dSph galaxies in the outer halo (Font et al. 2006).

7 Conclusions

The last decade has seen great progress in determinations of abundance ratios that can be used as tracers of stellar populations. This has been possible due to the availability of high-resolution echelle spectrographs at large telescopes – in some cases with multiplex capabilities so that many stars, e.g., in clusters and dwarf galaxies, can be observed simultaneously. Determinations of $[\alpha/\text{Fe}]$ from low- or medium-resolution spectroscopy, e.g., in the SDSS and RAVE surveys, are also important in connection with large statistical investigations of stellar populations. Similar measurements of $[\alpha/\text{Fe}]$ will be obtained in connection with the GAIA mission. Furthermore, abundances of Galactic bulge stars are now been obtained from high-resolution infrared spectra, but the spectral coverage is small and there is no multiplex advantages. Hence, the number of bulge stars studied with infrared spectra is still limited. More efficient infrared echelle spectrographs with multiplex capabilities are needed to take advantage of the lower reddening in the infrared spectral region, when dealing with the bulge and the inner part of the disk.

Elemental abundance ratios are derived by the aid of a model atmosphere analysis of the available stellar spectra, as discussed in [Sect. 2](#). Most studies are still based on homogeneous model atmospheres and the assumption of LTE, but several works show that inhomogeneous models and deviations from LTE can change the derived abundances significantly. Such 3D and non-LTE modeling is difficult to carry out and in some cases the derived abundances are sensitive to uncertain hydrogen collision cross sections. Hence, the derived trends of abundance ratios as a function of metallicity can be quite uncertain. On the other hand, it is possible to determine precise differences of abundance ratios by a 1D LTE analysis of a sample of stars confined to small ranges in effective temperature, surface gravity, and metallicity. In this way, abundance ratios can be used to disentangle stellar populations. Care should, however, be taken when comparing different spectral types and luminosity classes, such as F and G dwarfs with K giants, because systematic errors may be important.

As discussed in [Sect. 3](#), the abundance ratios $[\alpha/\text{Fe}]$, $[\text{C}/\text{O}]$, $[\text{Na}/\text{Fe}]$, $[\text{Ni}/\text{Fe}]$, $[\text{Cu}/\text{Fe}]$, $[\text{Eu}/\text{Ba}]$, and $[\text{Ba}/\text{Y}]$ are of high interest as tracers of stellar populations. The usefulness of these ratios is related to the fact that the elements involved are produced in different types of stars. The nucleosynthesis is not well understood for all elements, but it seems that variations of the ratios from one population to the next can be explained in terms of different rates of star formation and chemical evolution. In general, one does not need to invoke variations in the initial mass function to explain the abundance ratios. A better understanding of the nucleosynthesis of the elements is, however, important in order to learn more about the formation and evolution of the various populations.

The power of abundance ratios as population tracers is evident from [Fig. 2-16](#), where $[\alpha/\text{Fe}]$ is plotted as a function of $[\text{Fe}/\text{H}]$ for seven different populations. $[\alpha/\text{Fe}]$ is defined as the average value of $[\text{Mg}/\text{Fe}]$, $[\text{Si}/\text{Fe}]$, $[\text{Ca}/\text{Fe}]$, and $[\text{Ti}/\text{Fe}]$ and may be used to estimate the timescale for the chemical enrichment of the population, as explained in [Sect. 3.2](#). On the basis of this figure and the more detailed discussion in the previous sections, the following conclusions about stellar populations in the Galactic disk, the bulge, and the halo can be made.

The disk consists of two main populations, the thin and the thick disk, which differ in $[\alpha/\text{Fe}]$ as well as $[\text{C}/\text{O}]$ and $[\text{Eu}/\text{Ba}]$. The thick disk is formed on a relatively short timescale with enrichment from type II SNe only up to $[\text{Fe}/\text{H}] \simeq -0.4$. Thin-disk stars have lower values of $[\alpha/\text{Fe}]$ due to type Ia SNe contributions to the chemical enrichment; hence, the timescale of evolution is longer than in the case of the thick disk. In the metallicity range $-0.7 < [\text{Fe}/\text{H}] < -0.4$, a gap in $[\alpha/\text{Fe}]$ is present between the two disks. Altogether, the $[\alpha/\text{Fe}]$ distribution of disk stars

is well explained by a scenario, for which a period of rapid star formation in the early Galactic disk was interrupted by a merging satellite galaxy that “heated” the already formed stars to thick-disk kinematics. This was followed by a hiatus in star formation, in which metal-poor gas was accreted and type Ia SNe caused $[\alpha/\text{Fe}]$ to decrease. When star formation resumed, the first thin-disk stars formed with low metallicity and low $[\alpha/\text{Fe}]$. However, as discussed in [Sect. 4.2](#), an alternative model with a monotonically decreasing star formation rate and radial migration of stars and gas (Schönrich and Binney 2009a, b) also predicts a bimodal distribution of $[\alpha/\text{Fe}]$. Very precise measurements of the distributions of $[\alpha/\text{Fe}]$ as well as $[\text{C}/\text{O}]$ and $[\text{Eu}/\text{Ba}]$ for a large volume-limited sample of F and G main-sequence stars would be important to distinguish between the two competing models for disk formation.

For the bulge, there has been great progress in studies of abundance ratios in recent years. In some works, enhanced values of $[\alpha/\text{Fe}] \sim +0.3$ are found for stars as metal-rich as the Sun, but in the most precise differential studies (Alves-Brito et al. 2010), the trend of $[\alpha/\text{Fe}]$ for the bulge is found to be similar to that of the thick disk ([Fig. 2-10](#)). The abundances derived by Bensby et al. (2011) for microlensed main-sequence stars ([Fig. 2-16](#)) suggest that the bulge may consist of two distinct populations: i.e., old metal-poor stars with enhanced $[\alpha/\text{Fe}]$ ratios related to the thick disk, and younger very metal-rich stars with disk-like α/Fe ratios. Still, one would like to see many more bulge stars observed before drawing conclusions concerning models for bulge formation from the abundance ratios.

For halo stars in the solar neighborhood, there is evidence for the existence of two distinct populations clearly separated in $[\alpha/\text{Fe}]$, $[\text{Na}/\text{Fe}]$, and $[\text{Ni}/\text{Fe}]$, as discussed in [Sect. 6.1](#). The high- α stars have abundance ratios very similar to thick-disk stars. They may be ancient stars formed in the Galactic disk or bulge and ejected to the halo by merging satellite galaxies (Purcell et al. 2010; Zolotov et al. 2009, 2010), or they may simply belong to the high-velocity tail of a thick disk with a non-Gaussian velocity distribution. The low- α stars tend to have retrograde motions, and many of them move on elongated radial orbits close to the Galactic plane as predicted for the stellar debris of the captured ω Cen progenitor galaxy. It is likely that the low-alpha stars have been accreted from dSph galaxies with a relatively slow chemical evolution, for which type Ia SNe started to contribute iron at $[\text{Fe}/\text{H}] \sim -1.5$. Perhaps future precise studies of abundance ratios of larger samples of halo stars will reveal additional subpopulations.

Stars in surviving dSph galaxies have even lower values of $[\alpha/\text{Fe}]$ than the low- α halo stars, as seen from [Fig. 2-16](#). This is to be expected according to simulations of the chemical evolution of a hierarchically formed stellar halo in a Λ CDM Universe (Font et al. 2006). Present-day dSph galaxies in the outer halo have experienced a slower chemical evolution than more massive satellite galaxies accreted in the early Galaxy.

All globular clusters seem to consist of multiple stellar populations characterized by different values of $[\text{O}/\text{Fe}]$, $[\text{Na}/\text{Fe}]$, $[\text{Al}/\text{Fe}]$, and $[\text{Mg}/\text{Fe}]$. This may be due to chemical enrichment from intermediate-mass AGB stars undergoing hot-bottom hydrogen burning. There is also evidence for variations of $[\text{Fe}/\text{H}]$ and the abundance of *s*-process elements in several globular clusters, most notable in ω Cen. In this connection, it has been suggested that globular clusters were once the nuclei of now dissolved dwarf galaxies.

Precise abundance ratios of field stars belonging to the Galactic halo have so far only been obtained in a small region around the Sun. It would be important to extend such studies to more distant halo regions. To do this in an efficient way, one needs a fiber-coupled high-resolution spectrograph that would make it possible to observe many stars simultaneously over a relatively large field, say 1° in diameter. Such a spectrograph would also be very useful in exploring abundance gradients in the Galactic disk, both radially and in the direction toward the Galactic poles.

Cross-References

- [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- [Dynamics of Disks and Warps](#)
- [Globular Cluster Dynamical Evolution](#)
- [Mass Distribution and Rotation Curve in the Galaxy](#)
- [Metal-Poor Stars and the Chemical Enrichment of the Universe](#)
- [Open Clusters and their Role in the Galaxy](#)
- [Star Counts and the Nature of the Galactic Thick Disk](#)
- [The Galactic Bulge](#)
- [The Stellar and Sub-Stellar Initial Mass Function of Simple and Composite Populations](#)

References

- Alves-Brito, A., Meléndez, J., Asplund, M., Ramírez, I., & Yong, D. 2010, *A&A*, 513, A35
- Arlandini, C., Käppeler, F., Wisshak, K., et al. 1999, *ApJ*, 525, 886
- Arnett, W. D. 1971, *ApJ*, 166, 153
- Asplund, M. 2005, *ARA&A*, 43, 481
- Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, *ARA&A*, 47, 481
- Barklem, P. S., Christlieb, N., Beers, T. C., et al. 2005, *A&A*, 439, 129
- Bekki, K., & Freeman, K. C. 2003, *MNRAS*, 346, L11
- Bensby, T., & Feltzing, S. 2006, *MNRAS*, 367, 1181
- Bensby, T., Feltzing, S., Lundström, I., & Ilyin, I. 2005, *A&A*, 433, 185
- Bensby, T., Adén, D., Meléndez, J., et al. 2011, *A&A*, 533, A134
- Bergemann, M., & Cescutti, G. 2010, *A&A*, 522, A9
- Bergemann, M., & Gehren, T. 2008, *A&A*, 492, 823
- Bergemann, M., Pickering, J. C., & Gehren, T. 2010, *MNRAS*, 401, 1334
- Boeche, C., Siebert, A., & Steinmetz, M. 2008, *AIP Conf. Ser.*, 1082, 61
- Busso, M., Gallino, R., & Wasserburg, G. J. 1999, *ARA&A*, 37, 239
- Carollo, D., Beers, T. C., Lee, Y. S., et al. 2007, *Nature*, 450, 1020
- Carretta, E., Bragaglia, A., Gratton, R. G., & Lucatello, S. 2009, *A&A*, 505, 139
- Carretta, E., Bragaglia, A., Gratton, R., et al. 2010, *ApJ*, 712, L21
- Casagrande, L., Ramírez, I., Meléndez, J., Bessell, M., & Asplund, M. 2010, *A&A*, 512, A54
- Cayrel, R., Depagne, E., Spite, M., et al. 2004, *A&A*, 416, 1117
- Chen, Y. Q., Nissen, P. E., Zhao, G., Zhang, H. W., & Benoni, T. 2000, *A&AS*, 141, 491
- Chen, Y. Q., Nissen, P. E., Zhao, G., & Asplund, A. 2002, *A&A*, 390, 225
- Chiappini, C., Hirschi, R., Meynet, G., et al. 2006, *A&A*, 449, L27
- Cooke, R., Pettini, M., Steidel, C. C., Rudie, G. C., & Nissen, P. E. 2011, *MNRAS*, 417, 1534
- Cunha, K., Smith, V. V., Suntzeff, N. B., et al. 2002, *AJ*, 124, 379
- Decressin, T., Meynet, G., Charbonnel, C., Prantzos, N., & Ekström, S. 2007, *A&A*, 464, 1029
- D'Ercole, A., Vesperini, E., D'Antona, F., et al. 2008, *MNRAS*, 391, 825
- Drawin, H. W. 1969, *Z. Phys.*, 225, 483
- Edvardsson, B., Andersen, J., Gustafsson, B., et al. 1993, *A&A*, 275, 101
- Eggen, O. J., Lynden-Bell, D., & Sandage, A. R. 1962, *ApJ*, 136, 748
- Elmegreen, B. G., Bournaud, F., & Elmegreen, D. M. 2008, *ApJ*, 688, 67
- Fabbian, D., Nissen, P. E., Asplund, M., Pettini, M., & Akerman, C. 2009, *A&A*, 500, 1143
- Fenner, Y., Gibson, B. K., Gallino, R., & Lugaro, M. 2006, *ApJ*, 646, 184
- Font, A. S., Johnston, K. V., Bullock, J. S., & Robertson, B. E. 2006, *ApJ*, 638, 585
- Fuhrmann, K. 2004, *AN*, 325, 3
- Fulbright, J. P. 2002, *AJ*, 123, 404
- Fulbright, J. P., McWilliam, A., & Rich, R. M. 2007, *ApJ*, 661, 1152
- Gehren, T., Liang, Y. C., Shi, J. R., Zhang, H. W., & Zhao, G. 2004, *A&A*, 413, 1045
- Gilmore, G., & Reid, N. 1983, *MNRAS*, 202, 1025
- González Hernández, J. I., & Bonifacio, P. 2009, *A&A*, 497, 497
- Gratton, R., Carretta, E., Matteucci, F., & Sneden, C. 1996, *ASP Conf. Ser.*, 92, 307
- Gratton, R. G., Bonifacio, P., Bragaglia, A., et al. 2001, *A&A*, 369, 87

- Gratton, R. G., Carretta, E., Desidera, S., et al. 2003, *A&A*, 406, 131
- Gratton, R. G., Sneden, S., & Carretta, E. 2004, *ARA&A*, 42, 385
- Gustafsson, B., Edvardsson, B., Eriksson, K., et al. 2008, *A&A*, 486, 951
- Haywood, M. 2008, *MNRAS*, 388, 1175
- Hill, V., Lecureur, A., Gómez, A., et al. 2011, *A&A*, 534, A80
- Ibata, R. A., Gilmore, G., & Irwin, M. J. 1994, *Nature*, 370, 194
- Israelian, G., & Rebolo, R. 2001, *ApJ*, 557, L43
- Kirby, E. N., Guhathakurta, P., Bolte, M., Sneden, C., & Geha, M. C. 2009, *ApJ*, 705, 328
- Kormendy, J., & Kennicutt, R. C., Jr., 2004, *ARA&A*, 42, 603
- Korn, A. J., Grundahl, F., Richard, O., et al. 2007, *ApJ*, 671, 402
- Kurucz, R. 1993, *ATLAS9 Stellar Atmosphere Programs and 2 km/s Grid*. Kurucz CD-ROM No. 13, Cambridge, Mass., Smithsonian Astrophysical Observatory
- Lecureur, A., Hill, V., Zoccali, M., et al. 2007, *A&A*, 465, 799
- Lee, J-W., Kang, Y-W., Lee, J., & Lee, Y-W. 2009, *Nature*, 462, 480
- Lee, Y. S., Beers, T. C., An, D., et al. 2011, *ApJ*, 738, 187
- Luck, R. E., Kovtyukh, V. V., & Andrievsky, S. M. 2006, *AJ*, 132, 902
- Marino, A. F., Milone, A. P., Piotto, G., et al. 2009, *A&A*, 505, 1099
- Mashonkina, L., Gehren, T., Travaglio, C., & Borkova, T. 2003, *A&A*, 397, 275
- Mashonkina, L., Gehren, T., Shi, J. -R., Korn, A. J., & Grupp, F. 2011, *A&A*, 528, A87
- McWilliam, A., & Rich, R. M. 1994, *ApJS*, 91, 749
- McWilliam, A., Preston, G. W., Sneden, C., & Searle, L. 1995, *AJ*, 109, 2757
- Meléndez, J., Asplund, M., Alves-Brito, A., et al. 2008, *A&A*, 484, L21
- Meléndez, J., Asplund, M., Gustafsson, B., & Yong, D. 2009, *ApJ*, 704, L66
- Meza, A., Navarro, J. F., Abadi, M. G., & Steinmetz, M. 2005, *MNRAS*, 359, 93
- Mishenina, T. V., Kovtyukh, V. V., Soubiran, C., Travaglio, C., & Busso, M. 2002, *A&A*, 396, 189
- Neves, V., Santos, N. C., Sousa, S. G., Correia, A. C. M., & Israelian, G. 2009, *A&A*, 497, 563
- Nissen, P. E. 2008, *Phys. Scr.*, T133, 014022
- Nissen, P. E., & Schuster, W. J. 1997, *A&A*, 326, 751
- Nissen, P. E., & Schuster, W. J. 2010, *A&A*, 511, L10
- Nissen, P. E., & Schuster, W. J. 2011, *A&A*, 530, A15
- Nissen, P. E., Primas, F., Asplund, M., & Lambert, D. L. 2002, *A&A*, 390, 235
- Nissen, P. E., Akerman, C., Asplund, M., et al. 2007, *A&A*, 469, 319
- Nordström, B., Mayor, M., Andersen, J., et al. 2004, *A&A*, 418, 989
- Pompéia, L., Hill, V., Spite, M., et al. 2008, *A&A*, 480, 379
- Purcell, C. W., Bullock, J. S., & Kazantzidis, S. 2010, *MNRAS*, 404, 1711
- Ramírez, I., Allende Prieto, C., & Lambert, D. L. 2007, *A&A*, 465, 271
- Reddy, B. E., Tomkin, J., Lambert, D. L., & Allende Prieto, C. 2003, *MNRAS*, 340, 304
- Reddy, B. E., Lambert, D. L., & Allende Prieto, C. 2006, *MNRAS*, 367, 1329
- Romano, D., & Matteucci, F. 2007, *MNRAS*, 378, L59
- Ryde, N., Gustafsson, B., Edvardsson, B., et al. 2010, *A&A*, 509, A20
- Sbordone, L., Bonifacio, P., Buonanno, R., et al. 2007, *A&A*, 465, 815
- Schönrich, R., & Binney, J. 2009a, *MNRAS*, 396, 203
- Schönrich, R., & Binney, J. 2009b, *MNRAS*, 399, 1145
- Schuster, W. J., Moitinho, A., Márquez, A., Parrao, L., & Covarrubias, E. 2006, *A&A*, 445, 939
- Searle, L., & Zinn, R. 1978, *ApJ*, 225, 357
- Smith, V. V., Suntzeff, N. B., Cunha, K., et al. 2000, *AJ*, 119, 1239
- Sousa, S. G., Santos, N. C., Israelian, G., Mayor, M., & Monteiro, M. J. P. F. G. 2007, *A&A*, 469, 783
- Stephens, A., & Boesgaard, A. M. 2002, *AJ*, 123, 1647
- Strömgren, B. 1987, in *The Galaxy*, ed. G. Gilmore, & B. Carswell (Dordrecht: Reidel), 229
- Takeda, Y., Hashimoto, O., & Taguchi, H. 2005, *PASJ*, 57, 751
- Tolstoy, E., Hill, V., & Tosi, M. 2009, *ARA&A*, 47, 371
- Tsujimoto, T., Nomoto, K., Yoshii, Y., et al. 1995, *MNRAS*, 277, 945
- Turcotte, S., & Wimmer-Schweingruber, R. F. 2002, *J. Geophys. Res.*, 107, 1442
- Venn, K. A., Irwin, M., Shetrone, M. D., et al. 2004, *AJ*, 128, 1177
- Ventura, P., D'Antona, F., Mazzitelli, I., & Gratton, R. 2001, *ApJ*, 550, L65
- Yong, D., & Grundahl, F. 2008, *ApJ*, 672, L29
- Yong, D., Grundahl, F., Lambert, D. L., Nissen, P. E., & Shetrone, M. D. 2003, *A&A*, 402, 985
- Yong, D., Grundahl, F., Nissen, P. E., Jensen, H. R., & Lambert, D. L. 2005, *A&A*, 438, 875
- Yong, D., Carney, B. W., Teixeira de Almeida, M. L., & Pohl, B. L. 2006, *AJ*, 131, 2256
- Zoccali, M., Hill, V., Lecureur, A., et al. 2008, *A&A*, 486, 177
- Zolotov, A., Willman, B., Brooks, A. M., et al. 2009, *ApJ*, 702, 1058
- Zolotov, A., Willman, B., Brooks, A. M., et al. 2010, *ApJ*, 721, 738

3 Metal-Poor Stars and the Chemical Enrichment of the Universe

Anna Frebel¹ · John E. Norris²

¹Department of Physics, Massachusetts Institute of Technology & Kavli Institute for Astrophysics and Space Research, Cambridge, MA, USA

²Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT, Australia

1	<i>Introduction</i>	57
1.1	The Role of Metal-Poor Stars	58
1.2	Background Matters	59
1.2.1	Essential Reading	59
1.2.2	Abundance Definitions	59
1.2.3	Nomenclature	59
1.3	Plan of Attack	60
2	<i>Discovery: The Search for Needles in the Haystack</i>	61
2.1	Historical Perspective	61
2.2	Search Techniques	63
2.3	High-Resolution and High S/N Follow-Up Spectroscopy	64
2.4	Census of the Most Metal-Poor Stars	65
2.5	The Lowest Observable Metallicity	68
3	<i>Derived Chemical Abundances</i>	69
3.1	Abundance Determination	69
3.1.1	One-Dimensional Model Atmosphere Analyses	69
3.1.2	Three-Dimensional Model Atmospheres	70
3.1.3	Departures from Thermodynamic Equilibrium (Non-LTE)	70
3.1.4	Caveat Emptor	71
3.1.5	Post-Astration Abundance Modification	71
3.2	Abundance Patterns	72
3.2.1	Metallicity Distribution Functions (MDF)	72
3.2.2	Relative Abundances	75
4	<i>The Chemical Evolution of the Universe</i>	79
4.1	Relics of the Big Bang	79
4.1.1	Helium	79
4.1.2	Lithium	79

4.2	The Milky Way Halo	81
4.2.1	The Evolution of Carbon Through Zinc	82
4.2.2	The Evolution of Neutron-Capture Elements	87
4.3	The Milky Way Globular Clusters and Dwarf Galaxies	94
4.3.1	Globular Clusters	94
4.3.2	Dwarf Galaxies	94
5	<i>Cosmo-Chronometry</i>	96
5.1	Nucleo-chronometry of Metal-Poor Field Stars	98
6	<i>Cosmogony</i>	100
6.1	The Early Universe	100
6.2	The Milky Way	105
7	<i>Conclusions and Future Prospects</i>	108
	<i>Acknowledgements</i>	110
	<i>References</i>	111

Abstract: Metal-poor stars hold the key to our understanding of the origin of the elements and the chemical evolution of the Universe. This chapter describes the process of discovery of these rare stars, the manner in which their surface abundances (produced in supernovae and other evolved stars) are determined from the analysis of their spectra, and the interpretation of their abundance patterns to elucidate questions of origin and evolution.

More generally, studies of these stars contribute to other fundamental areas that include nuclear astrophysics, conditions at the earliest times, the nature of the first stars, and the formation and evolution of galaxies – including our own Milky Way. This is illustrated with results from studies of lithium formed during the Big Bang; of stars dated to within ~ 1 Gyr of that event; of the most metal-poor stars, with abundance signatures very different from all other stars; and of the buildup of the elements over the first several Gyr. The combination of abundance and kinematic signatures constrains how the Milky Way formed, while recent discoveries of extremely metal-poor stars in the Milky Way’s dwarf galaxy satellites constrain the hierarchical build-up of its stellar halo from small dark-matter dominated systems.

Two areas needing priority consideration are discussed. The first is improvement of abundance analysis techniques. While one-dimensional, local thermodynamic equilibrium (1D/LTE) model atmospheres provide a mature and precise formalism, proponents of more physically realistic 3D/non-LTE techniques argue that 1D/LTE results are not accurate, with systematic errors often of order ~ 0.5 dex or even more in some cases. Self-consistent 3D/non-LTE analysis as a standard tool is essential for meaningful comparison between the abundances of metal-poor stars and models of chemical enrichment.

The second need is for larger samples of metal-poor stars, in particular those with $[\text{Fe}/\text{H}] < -4$ and those at large distances (20–50 kpc), including the Galaxy’s ultra-faint dwarf satellites. With future astronomical surveys and facilities, these endeavors will become possible. This will provide new insights into small-scale details of nucleosynthesis as well as large-scale issues such as galactic formation.

Keywords: Abundances, Early Universe, Galaxy: formation, Galaxy: halo, Nuclear reactions, Nucleosynthesis, Stars: abundances

1 Introduction

A few minutes after the beginning of the Universe, the only chemical elements that existed were hydrogen (~ 0.75 by mass fraction), helium (~ 0.25), and a miniscule amount of lithium ($\sim 2 \times 10^{-9}$). Today, some 13.7 Gyr later, the mass fraction of the elements Li–U in the Milky Way Galaxy stands at ~ 0.02 , essentially all of it created by stellar nucleosynthesis. Metal-poor stars provide the foundation for our understanding of the intricate details of the way in which this enrichment occurred.

The astronomer Carl Sagan summarized cosmic chemical evolution in just one sentence, “We are made from star stuff.” Studying stars that are extremely underabundant in their heavy elements (collectively referred to as “metals”) takes us right to the heart of this statement.

These objects allow us to study the origin of the elements that were subsequently recycled in stellar generations over billions of years until ending up in the human body.

The rationale for analyzing metal-poor stars is that they are long-lived, low-mass objects, the majority of which are main sequence and giant stars that have preserved in their atmospheres the chemical signatures of the gas from which they formed. Given that the overall Universe was largely devoid of metals at the earliest times, it is generally assumed (and borne out by analysis) that low metallicity indicates old age. For these objects to be still observable, their masses are of order $0.6\text{--}0.8 M_{\odot}$. By measuring their surface composition today, one can “look back” in time and learn about the nature of the early Universe. Another vital assumption is that the stellar surface composition has not been significantly altered by any internal “mixing” processes or by external influences such as accretion of interstellar material that would change the original surface abundance.

Analysis of old, metal-poor stars to study the early Universe is often referred to as “stellar archaeology” and “near-field cosmology.” This fossil record of local Galactic metal-poor stars provides unique insight into the enrichment of the Universe, complementing direct studies of high-redshift galaxies.

1.1 The Role of Metal-Poor Stars

The abundances of the elements in stars more metal-poor than the Sun have the potential to inform our understanding of conditions from the beginning of time – the Big Bang – through the formation of the first stars and galaxies and up to the relatively recent time when the Sun formed. An incomplete list of the rationale for studying metal-poor stars includes the following:

- The most metal-poor stars ($[\text{Fe}/\text{H}] \lesssim -4.0$), with primitive abundances of the heavy elements (atomic number $Z > 3$), are most likely the oldest stars so far observed.
- The lithium abundances of extremely metal-poor near-main-sequence-turnoff stars have the potential to directly constrain conditions of the Big Bang.
- The most metal-poor objects were formed at epochs corresponding to redshifts $z > 6$, and probe conditions when the first heavy element-producing objects formed. The study of objects with $[\text{Fe}/\text{H}] < -3.5$ permits insight into conditions at the earliest times that is not readily afforded by the study of objects at high redshift.
- They constrain our understanding of the nature of the first stars, the initial mass function, the explosion of super- and hyper-novae, and how their ejecta were incorporated into subsequent early generations of stars.
- Comparison of detailed observed abundance patterns with the results of stellar evolution calculations and models of galactic chemical enrichment strongly constrains the physics of the formation and evolution of stars and their host galaxies.
- In some stars with $[\text{Fe}/\text{H}] \sim -3.0$, the overabundances of the heavy-neutron-capture elements are so large that a measurement of Th and U becomes possible which leads to independent estimates of their ages and hence of the Galaxy.
- Stars with $[\text{Fe}/\text{H}] \lesssim -0.5$ inform our understanding of the evolution of the Milky Way system. Relationships between abundance, kinematic, and age distributions – the defining characteristics of stellar populations – permit choices between the various paradigms of how the system formed and has evolved.

1.2 Background Matters

1.2.1 Essential Reading

The study of metal-poor stars for insight into the chemical evolution of the Universe has resulted in a rich literature, embracing diverse areas. The reader will find the following topics and reviews of considerable interest.

For the context of the early chemical enrichment of the Universe, and how one might use metal-poor stars to explore back in time to the Big Bang, see Bromm and Larson (2004), Frebel (2010), and Pagel (1997). To understand how one determines the chemical abundances of stars, the important abundance patterns, and how reliable the results are, we refer the reader to Wheeler et al. (1989), Sneden et al. (2008), and Asplund (2005). Other relevant questions and reviews include the following: How does one discover metal-poor stars: Beers and Christlieb (2005). What is the role of abundance in the stellar population paradigm: Sandage (1986), Gilmore et al. (1989), and Freeman and Bland-Hawthorn (2002). How do the abundances constrain galactic chemical enrichment: McWilliam (1997). What progress has been made in understanding the supernovae and hypernovae that produce the chemical elements: Timmes et al. (1995), Arnett (1996), and Kobayashi et al. (2006), and references therein. These reviews are of course not one-dimensional, and in many cases, they describe matters in several of the topics highlighted above. They will repay close reading by the interested student.

1.2.2 Abundance Definitions

Most basically, $\epsilon(A)$, the abundance of element A is presented logarithmically, relative to that of hydrogen (H), in terms of N_A and N_H , the numbers of atoms of A and H.

$$\log_{10} \epsilon(A) = \log_{10} (N_A/N_H) + 12$$

(For lithium, the abundance is mostly expressed as $A(\text{Li}) = \log \epsilon(\text{Li})$, and for hydrogen, by definition, $\log_{10} \epsilon(\text{H}) = 12$.) For stellar abundances in the literature, results are generally presented relative to their values in the Sun, using the so-called “bracket notation,”

$$\begin{aligned} [A/H] &= \log_{10} (N_A/N_H)_* - \log_{10} (N_A/N_H)_\odot \\ &= \log_{10} \epsilon(A)_* - \log_{10} \epsilon(A)_\odot, \end{aligned}$$

and for two elements A and B, one then has

$$[A/B] = \log_{10} (N_A/N_B)_* - \log_{10} (N_A/N_B)_\odot$$

In the case of the Fe metallicity, $[\text{Fe}/\text{H}] = \log_{10} (N_{\text{Fe}}/N_{\text{H}})_* - \log_{10} (N_{\text{Fe}}/N_{\text{H}})_\odot$. For example, $[\text{Fe}/\text{H}] = -4.0$ corresponds to an iron abundance 1/10,000 that of the Sun.

For completeness, it should be noted that with the bracket notation, one needs to know the abundance not only of the star being analyzed, but also of the Sun, the chemical composition of which has recently been revised substantially for some elements (Asplund et al. 2009).

1.2.3 Nomenclature

Baade (1944), in his seminal paper on the subject, defined two groups of stars, Type I and Type II, which today are referred to as Population I and Population II. The first referred to young

stars, including open clusters, which reside in the disk of the Galaxy, while the second includes its globular clusters and essentially all of its known metal-poor stars. In what follows, Population II will be referred to as the “halo,” which defines the spatial distribution of the population. It has been speculated that a so-called Population III exists, which comprises the elusive first stars. With the advent of detailed cosmological simulations of primordial star formation, the term “Population III” is now widely used only for stars that first formed from zero-metallicity gas that consisted only of hydrogen, helium, and traces of lithium. The most metal-poor stars currently known are thus extreme members of Population II.

Following Beers and Christlieb (2005) (with some modifications and additions), the nomenclature listed in [Table 3-1](#) will be adopted for different types of metal-poor stars in terms of population, metallicity, and chemical signatures. As can be seen, the main metallicity indicator is the iron abundance, $[Fe/H]$. Iron has the advantage that among the elements, it has the richest absorption line spectrum in the optical region, facilitating determination of Fe abundance independent of the wavelength range covered by the spectrum. With few exceptions, $[Fe/H]$ traces the overall metallicity of the stars fairly well.

1.3 Plan of Attack

For convenience, and the purposes of this chapter, the term “metal-poor” will be taken to mean stars in the Milky Way system having $[Fe/H] < -1.0$. This embraces all of the “metal-poor” categories of Beers and Christlieb (2005) shown in [Table 3-1](#). It will confine our attention

Table 3-1

Metal-poor star related definitions

Description	Definition	Abbreviation ^a
Population III stars	Postulated first stars, formed from zero-metallicity gas	Pop III
Population II stars	Old (halo) stars formed from low-metallicity gas	Pop II
Population I stars	Young (disk) metal-rich stars	Pop I
Solar	$[Fe/H] = 0.0$	
Metal-poor	$[Fe/H] < -1.0$	MP
Very metal-poor	$[Fe/H] < -2.0$	VMP
Extremely metal-poor	$[Fe/H] < -3.0$	EMP
Ultra metal-poor	$[Fe/H] < -4.0$	UMP
Hyper metal-poor	$[Fe/H] < -5.0$	HMP
Carbon-rich stars	$[C/Fe] > +0.7$ for $\log(L/L_{\odot}) \leq 2.3$	CEMP
	$[C/Fe] \geq (+3.0 - \log(L/L_{\odot}))$ for $\log(L/L_{\odot}) > 2.3$	CEMP
n-capture-rich stars	$0.3 \leq [Eu/Fe] \leq +1.0$ and $[Ba/Eu] < 0$	r-I
n-capture-rich stars	$[Eu/Fe] > +1.0$ and $[Ba/Eu] < 0$	r-II
n-capture-rich stars	$[Ba/Fe] > +1.0$ and $[Ba/Eu] > +0.5$	s
n-capture-rich stars	$0.0 < [Ba/Eu] < +0.5$	r/s
n-capture-normal stars	$[Ba/Fe] < 0$	no

Note – Carbon-rich stars appear with r- and s-process enhancements also. The CEMP definitions are from Aoki et al. (2007) and differ somewhat from Beers and Christlieb (2005)

^aCommonly used in the literature

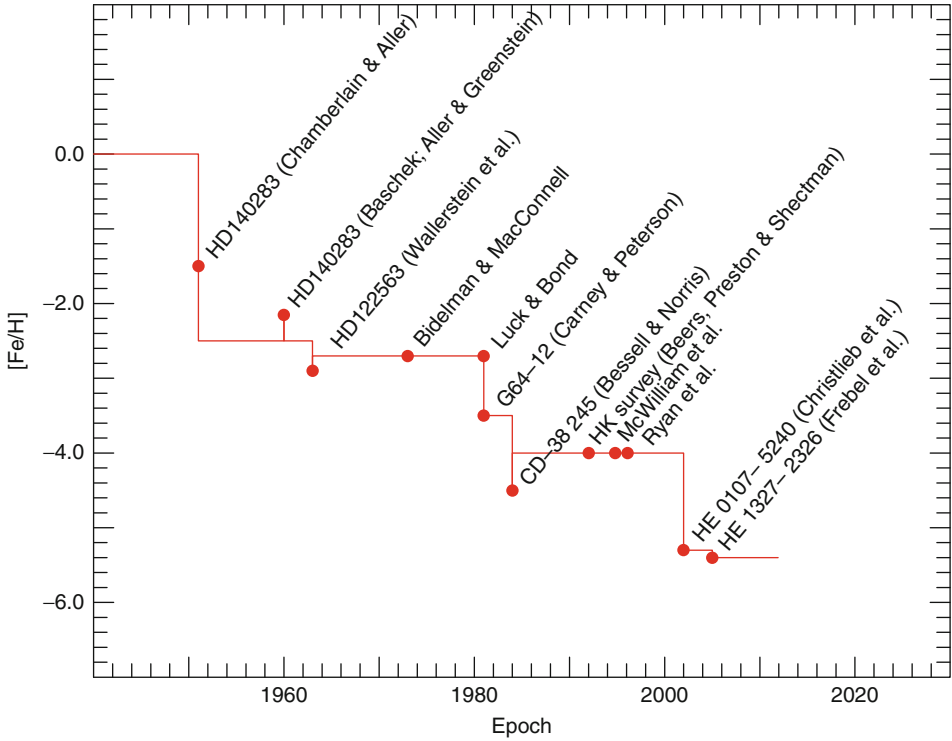
principally to field stars and globular clusters of the galactic halo and the Galaxy's dwarf galaxy satellites. Further, if one accepts the Galactic age-metallicity relationship presented, for example, by Freeman and Bland-Hawthorn (2002), this restricts discussion to star formation and associated chemical enrichment that occurred during the first ~ 4 Gyr following the Big Bang. Our abundance restriction also includes part of the so-called metal-weak thick disk (MWTD) (see Chiba and Beers 2000) and the Galactic bulge, neither of which will be discussed here. Currently, no major works have been carried out that attempt to elucidate differences between halo and MWTD abundance patterns. The bulge, on the other hand, is believed to be the site of some of the very first star formation, the result of which is seen today admixed with overwhelming later generations at the Galaxy's center, precluding insight into its metal-poor population. This is work for the future.

In [Sect. 2](#), the search for metal-poor stars will be briefly outlined, and what has been discovered so far. [Section 3](#) is concerned with the manner in which chemical abundances are determined (and their reliability). An overview of metallicity distribution functions (MDF) of globular clusters, field stars, and satellite dwarf galaxies is also given. A major focus of this chapter is an introduction to the interpretation of the relative abundances, $[X/Fe]$, and the corresponding chemical patterns observed in metal-poor stars. Against this background, in [Sect. 4](#), the body of metal-poor stellar abundances is presented, and the general abundance trends are discussed in light of expectations set by models of stellar evolution and galactic chemical evolution (GCE). In [Sect. 5](#), age determination in a small class of extremely metal-poor stars which have huge r-process element enhancement is described. In [Sect. 6](#), the implications of deduced abundances for the cosmogony of the early Universe and the Milky Way system are considered. Finally, in [Sect. 7](#), the possibilities and challenges of the future are outlined.

2 Discovery: The Search for Needles in the Haystack

2.1 Historical Perspective

Chemical abundance (along with spatial distribution, kinematics, and age) is one of the basic parameters that define stellar populations (see, e.g., Sandage 1986). In the middle of the twentieth century, however, as noted by Sandage, "There had grown up a general opinion, shared by nearly all spectroscopists, that there was a single universal curve of all the elements, and that the Sun and all the stars shared ...precisely ...this property of identical ratios of any element A to hydrogen." Subtle spectroscopic differences that had been documented at that time were thought to result from differences in physical conditions in the atmospheres of stars rather than in chemical composition. Chamberlain and Aller (1951) profoundly changed this concept with their chemical abundance analysis of the "A-type subdwarfs" HD 19445 and HD 140243, for which they reported $[Fe/H]$ ($[Ca/H]$) = -0.8 (-1.4) and -1.0 (-1.6), respectively. Their work clearly established the existence of stars with elemental abundances (relative to that of hydrogen) lower than in the Sun and that these lower abundances played a critical role in determining the strength of their spectral features. (It should be noted in passing that these early values are believed to have been overestimates, with the currently accepted values for Fe being $[Fe/H] \sim -2.0$ and -2.5 , respectively. See Sandage (1986, Footnote 2) for an interesting sociological comment on the differences between the earlier and current values.) Soon



■ Fig. 3-1

[Fe/H] for the most metal-poor star then known as a function of epoch. The *symbols* denote the abundance determined by the authors, while the *horizontal lines* refer, approximately, to currently accepted values (The abundances are based on one-dimensional, local thermodynamic equilibrium model atmosphere analysis. See ▶ Sect. 3.1)

after that work, Burbidge et al. (1957) reviewed the case for the nucleosynthesis of almost all of the chemical elements within stars. In the decades that followed, exhaustive searches for, and analysis of “metal-poor” stars – as illustrated in ▶ Fig. 3-1 – have led to the discovery of stars with lower and lower values of [Fe/H] until, at time of writing, two objects with [Fe/H] ~ -5.5 are known.

Two major developments, relevant to the present discussion, occurred in parallel with the early chemical abundance analyses of stars. The first was the wide acceptance of the “Big Bang” paradigm as the most likely description of the Universe. The second was the demonstration by Wagoner et al. (1967), for example, that at the era of decoupling of radiation and matter, some minutes after the singularity, no elements beyond lithium had been produced (if isotropy and homogeneity were assumed).

One might then enquire how best observationally to examine the way in which the chemical enrichment of the Universe proceeded. A first approach would be to investigate objects at high redshift, such as galaxies and the Lyman- α clouds seen in the spectra of quasars. Songaila (2001) and Ryan-Weber et al. (2009) report that for redshifts $z = 5.0$ and 6.0 , measures of Si IV and C IV

(Songaila) and C IV (Ryan-Weber et al.) observed column densities imply intergalactic metallicity $Z_{\text{IGM}} \gtrsim 3.5 \times 10^{-4} Z_{\odot}$ and $(9 \pm 5) \times 10^{-5} Z_{\odot}$, respectively. Assuming solar abundance ratios, these intergalactic values correspond to $[\text{Fe}/\text{H}] \gtrsim -3.4$ and -4.0 . It is also important to note in this context the recent analyses of very metal-poor Damped Lyman- α systems (see Cooke et al. 2011) that are currently observed out to redshifts $z \sim 2-3$, and which report abundances of ~ 6 elements, down to $[\text{Fe}/\text{H}] \sim -3.0$. Far-field cosmological measurements thus currently reach to abundance limits 30 times larger than those observed in the most metal-poor stars in the Milky Way. Further, while to date only C and Si are observed at high redshift ($z \gtrsim 5$), some eight to nine elements are measurable in Galactic stars observed to have $[\text{Fe}/\text{H}] \sim -5.5$ (Christlieb et al. 2002; Frebel et al. 2005). That is to say, it seems reasonable to suggest that the most metal-poor stars have the potential to serve as the best cosmological probes of chemical enrichment at the earliest times.

2.2 Search Techniques

Metal-poor field stars are rare. To begin with, the proportion of stars in the solar neighborhood that belong to the halo population is only $\sim 10^{-3}$ (see, e.g., Bahcall and Soneira 1980). Further, as a rule of thumb, the simple chemical enrichment model of the halo of Hartwick (1976; see [Sect. 3.2.1](#) below) suggests that the number of stars should decrease by a factor of 10 for each factor of 10 decrease in abundance. For example, the number of stars with $[\text{Fe}/\text{H}] < -3.5$ should be smaller by a factor 100 than the number with $[\text{Fe}/\text{H}] < -1.5$. (For observational support for this suggestion, down to $[\text{Fe}/\text{H}] \sim -4.0$, below which it breaks down, see Norris 1999.) Roughly speaking, given that the stellar halo MDF peaks at $[\text{Fe}/\text{H}] = -1.5$, in the solar neighborhood one might expect to find ~ 1 in 200,000 stars with $[\text{Fe}/\text{H}] < -3.5$.

One thus needs to filter out disk stars if one wishes to find metal-poor stars. While important bright extremely metal-poor stars have been discovered somewhat serendipitously (e.g., the red giant CD-38° 245 with $[\text{Fe}/\text{H}] = -4.0$ and $V = 12.8$; Bessell and Norris 1984), for stars brighter than $B \sim 16$, this has to date been systematically achieved in one of two ways. The first uses the fact that the halo does not share the rotation of the Galactic disk and that a large fraction of its members have relatively high proper motions. The first star with $[\text{Fe}/\text{H}] < -3.0$ (G64-12; Carney and Peterson 1981) was discovered in this way. The major surveys to date that utilized this technique are those of Ryan and Norris (1991a) and Carney et al. (1996), whose samples each comprise a few hundred halo main-sequence dwarfs with $[\text{Fe}/\text{H}] < -1.0$ and who together report ~ 10 stars having $[\text{Fe}/\text{H}] < -3.0$.

The second method has been more prolific and utilizes objective-prism spectroscopy with Schmidt telescopes, which permit one to simultaneously obtain low-resolution spectra (resolving power $R (= \lambda/\Delta\lambda) \sim 400$) of many stars over several square degrees. Examination of the strength of the Ca II K line at 3,933.6 Å with respect to that of nearby hydrogen lines or an estimate of the color of the star permits one to obtain a first estimate of whether the star is metal-weak or not. Candidate metal-poor stars are then observed at intermediate resolution ($R \sim 2,000$) to obtain a measurement of the metal abundance of the star. The techniques are described in detail by Beers and Christlieb (2005), who also document important surveys that have obtained first abundance estimates for some tens of thousands of stars brighter than $B \sim 16.5$ with $[\text{Fe}/\text{H}] < -1.0$.

The most important Schmidt surveys to date have been the HK survey (Beers et al. 1992) and the Hamburg/ESO survey (HES) (Christlieb et al. 2008). In order to give the reader an

appreciation of the scope and many steps involved in the process, here is a brief description of the HES. According to N. Christlieb, the HES consists of some 12 million stars in the magnitude range $10 < B < 18$. In an effective survey area of some $6,700 \text{ deg}^2$, $\sim 21,000$ candidate metal-poor stars were selected, for which, at the time of writing, follow-up spectroscopy has been obtained of $\sim 5,200$. Preliminary metal-poor candidates were selected in several steps to arrive at candidate lists for which medium-resolution spectroscopy was sought. Due to limitations of telescope time and target faintness, it was common that not all stars could be observed. In the original candidate list in the magnitude range $13.0 \lesssim B \lesssim 17.5$ there were $\sim 3,700$ red giants, of which about 1,700 were observed at medium-resolution (Schörck et al. 2009), together with $\sim 3,400$ near-main-sequence-turnoff stars, of which ~ 700 have follow-up spectroscopy (Li et al. 2010). There is also a bright sample of $\sim 1,800$ stars having $B < 14.5$, for all of which medium resolution spectra were obtained by Frebel et al. (2006). From these samples, the most metal-poor candidates were selected for high-resolution spectroscopic observation. Various considerations determined whether a star was ultimately observed. These include telescope time allocations, observability and weather conditions during observing time, target brightness, reliability of the medium-resolution result, science questions to be addressed, and of course the preliminary metallicity of the star. Given these limitations, fainter stars remain unobserved on the target lists due to time constraints.

To this point, the discussion has been confined to surveys that have concentrated on discovering candidate metal-poor stars with $B \lesssim 17.5$, with follow-up medium-resolution spectroscopy complete in most cases to only somewhat brighter limits. Surveys that reach to considerably fainter limits are the Sloan Digital Sky Survey (SDSS) and the subsequent Segue-I and II surveys (see <http://www.sdss.org>), which have obtained spectra with resolving power $R \sim 2,000$ and are also proving to be a prolific source of metal-poor stars. In a sample of some 400,000 stars, SDSS/Segue has discovered 26,000 stars with spectra having $S/N > 10$ and $[\text{Fe}/\text{H}] < -2.0$ (based on these intermediate-resolution spectra), while some 400 have $[\text{Fe}/\text{H}] < -3.0$.

The search for metal-poor stars remains a very active field, with several exciting projects coming to completion, currently in progress, and planned. This matter will be further discussed in [Sect. 7](#).

2.3 High-Resolution and High S/N Follow-Up Spectroscopy

The final observational step in the discovery process is spectroscopy of the most significant objects (e.g., most metal-poor, or most chemically peculiar) at very high resolving power ($R \sim 10^4 - 10^5$) and $S/N \gtrsim 100$, in order to reveal the fine detail required for the determination of parameters such as accurate chemical abundances, isotope ratios, and in some cases stellar ages. This is best achieved with 6–10-m telescope/échelle spectrograph combinations – currently HET/HRS, Keck/HIRES, Magellan/MIKE, Subaru/HDS, and VLT/UVES.

In order to give the reader a feeling for the effect that increasing resolution and decreasing metallicity have on the observed flux, [Fig. 3-2](#) shows the increase in spectroscopic detail between intermediate ($R \sim 1,600$) and high ($R \sim 40,000$) resolving power for four metal-poor red giants of similar effective temperature (T_{eff}) and surface gravity ($\log g$) as metal abundance decreases from $[\text{Fe}/\text{H}] = -0.9$ to -5.4 (for HE 0107–5240, the most metal-poor giant currently known).

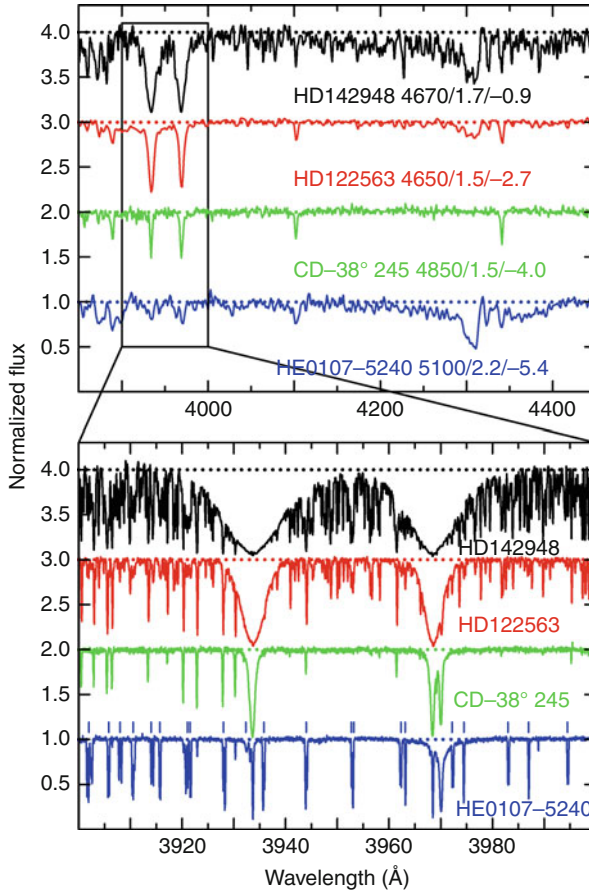
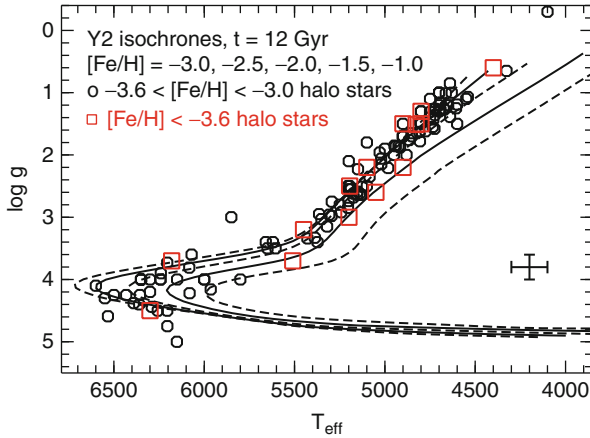


Fig. 3-2

(Upper panel) Spectra at intermediate resolution ($R \sim 1,600$) of metal-poor red giants over the range $-5.4 < [\text{Fe}/\text{H}] < -0.9$. Note the strong decrease in the strengths of the Ca II H & K lines at 3,933.6 and 3,968.4 Å. The numbers in the panel represent $T_{\text{eff}}/\log g/[\text{Fe}/\text{H}]$. (Lower panel) Spectra of the same stars at $R \sim 40,000$ on the range 3,900–4,000 Å. Note that while the Ca II H and K lines are very weak in the most metal-poor giant, HE 0107–5240, many more lines have appeared. These are features of CH (the positions of which are indicated immediately above the spectrum) resulting from an extremely large overabundance of carbon relative to iron in this object

2.4 Census of the Most Metal-Poor Stars

This section presents a census of stars having $[\text{Fe}/\text{H}] < -3.0$ and for which detailed high-resolution, high S/N , published abundance analyses are available. The data set comprises some 130 objects and may be found in the compilation of Frebel (2010). Figure 3-3 shows the distribution of the stellar parameters effective temperature, T_{eff} , and surface gravity, $\log g$,



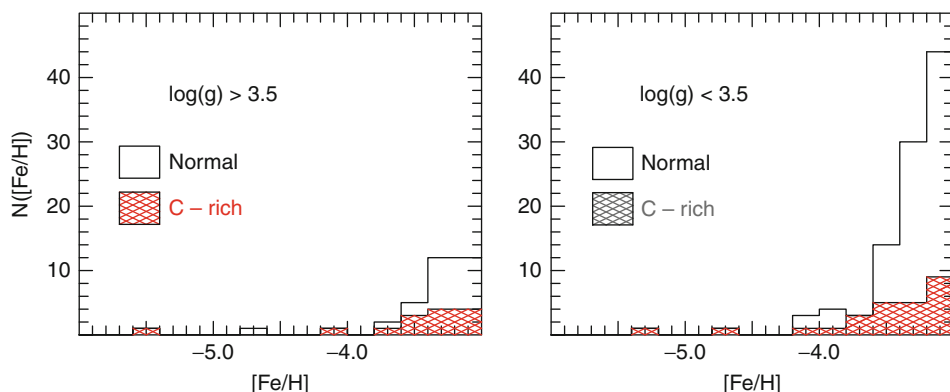
■ Fig. 3-3

Hertzsprung–Russell diagram of the ~ 130 known metal-poor stars with $[\text{Fe}/\text{H}] < -3.0$ studied at high-resolution (from the compilation of Frebel (2010)). Fourteen stars with $[\text{Fe}/\text{H}] < -3.6$ are marked with *open squares*. Typical error bars are indicated at *bottom right*. Several 12 Gyr isochrone tracks (from <http://www.astro.yale.edu/demarque/yyiso.html>) for different metallicities are overplotted for illustration. As can be seen, the main-sequence turnoff shifts significantly to hotter temperatures at $[\text{Fe}/\text{H}] < -2.0$, whereas the giant branch is less affected

of these objects, in comparison with several 12 Gyr isochrones of different metallicities, $[\text{Fe}/\text{H}]$, in the Hertzsprung–Russell diagram. The vast majority of the stars are luminous red giants ($4,000 \text{ K} < T_{\text{eff}} < 5,500 \text{ K}$, $0.0 < \log g < 3.5$), but about 25 main-sequence stars near the turnoff ($5,800 \text{ K} < T_{\text{eff}} < 6,600 \text{ K}$, $3.5 < \log g < 4.5$) are also known. A similar ratio is maintained for metallicities below $[\text{Fe}/\text{H}] = -3.6$. (The atmospheric parameters T_{eff} , $\log g$, and $[\text{Fe}/\text{H}]$ are the essential stellar parameters that determine the structure of a star’s outer layers and the details of its emergent flux, as will be discussed in [Sect. 3.1](#).)

► [Figure 3-4](#) presents the histogram of $[\text{Fe}/\text{H}]$ for the same group of objects. Two things are worth noting from this diagram. First, the number of stars decreases precipitously as one moves toward lowest abundance, and second, the proportion of carbon-rich objects increases dramatically. The implications of this remarkable behavior are discussed further in [Sect. 6.1](#). More generally, roughly 10% of the objects in this sample (often those with enhanced carbon) reveal a chemical nature that is different from that of “normal” halo stars. These objects are of particular interest as they can be used to address a variety of astrophysically important questions. Arguably, the most interesting stars in the sample are those with metallicities $[\text{Fe}/\text{H}] \lesssim -3.5$. It is probably reasonable to say that all stars known to have $[\text{Fe}/\text{H}] \lesssim -3.5$ from medium-resolution spectroscopy and $B < 16.5$ are included in this diagram, given their potential for insight into the early Universe.

There are eight stars known to have $[\text{Fe}/\text{H}] \sim -4.0$ or less. One of them is CD–38° 245 (Bessell and Norris 1984; Norris et al. 2001), the first star with $[\text{Fe}/\text{H}] \sim -4.0$. It was discovered some 30 years ago and was a long-standing record holder for the most iron-poor object in



■ Fig. 3-4

[Fe/H] histogram for stars having high-resolution, high S/N , abundance analyses and $[\text{Fe}/\text{H}] < -3.0$, from the compilation of Frebel (2010). On the left are results for main-sequence and subgiant stars, while the right presents data for red giants. The shaded regions refer to C-rich stars. Note the rapid decline in the number of stars as $[\text{Fe}/\text{H}]$ decreases, accompanied by an increase in the proportion of carbon-rich objects

the Milky Way. It has been observed and analyzed many times by most research groups working in the field. Only four stars in this very small sample have $[\text{Fe}/\text{H}] < -4.3$, with two having $[\text{Fe}/\text{H}] < -5.0$. In 2001, the first star with $[\text{Fe}/\text{H}] < -5.0$ was discovered. Until then, it had not been clear whether objects with metallicities lower than that of CD-38° 245 existed. This object, HE 0107-5240, is a faint ($V = 15.2$) red giant with $[\text{Fe}/\text{H}] = -5.3$ (Christlieb et al. 2002). In 2004, the bright ($V = 13.5$) subgiant HE 1327-2326 was identified and shown to have $[\text{Fe}/\text{H}] = -5.4$ (Frebel et al. 2005), corresponding to $\sim 1/250,000$ of the solar iron abundance. This small stellar Fe number density translates to an actual iron mass that is about 100 times less than that of the Earth's iron core! Both stars were found in the Hamburg/ESO survey. Since then, no further objects with such record-low Fe values have been discovered. As outlined in ▶ Sect. 7, new surveys will provide additional chances to uncover more of these rare stars.

The paucity of stars with $-5.3 \lesssim [\text{Fe}/\text{H}] \lesssim -4.3$ sparked considerable interest among theorists, with some suggesting that there may be a physical reason for this apparent gap in the metallicity distribution function (e.g., Shige-yama et al. 2003). The discovery, however, of the red giant HE 0557-4840 (Norris et al. 2007) and the dwarf star SDSS J102915+172927 (Caffau et al. 2011) both with $[\text{Fe}/\text{H}] \sim -4.7$ (adopting 1D Fe abundances) confirmed that extremely limited discovery statistics below $[\text{Fe}/\text{H}] \sim -4.3$, driven by only four stars, are most likely the cause of the apparent gap.

In summary, as of mid-2010, numerous stars with $[\text{Fe}/\text{H}] < -3.0$ have been discovered and many (~ 130) have been analyzed with high-resolution spectroscopy. Stars with $[\text{Fe}/\text{H}] < -3.5$ are much rarer, but most likely all known examples (~ 25) of them have high-resolution spectroscopic analyses. Only four stars with $[\text{Fe}/\text{H}] < -4.3$ are known, of which two have $[\text{Fe}/\text{H}] < -5.0$.

2.5 The Lowest Observable Metallicity

What is the lowest abundance one might be able to observe in the Galactic halo? From a practical point of view, a useful limit is set by the abundance corresponding to a measured Ca II K line strength of $20 \text{ m}\text{\AA}$ (roughly two to four times the strength on the weakest lines measurable in high-resolution, high S/N , spectra such as those shown in [Fig. 3-2](#)) in a cool red giant with $T_{\text{eff}} = 4,500 \text{ K}$ and $\log g = 1.5$. (Ca II K is the strongest atomic feature in metal-poor stars, and its strength is greater in red giants than near-main-sequence dwarfs due to the lower effective temperatures of the former.) Also, the abundances of red giants are much less modified by accretion from the Galactic interstellar medium (ISM) than those of main-sequence stars, because of the deep outer convective regions in giants.)

Adopting a 1D LTE model atmosphere (see [Sect. 3.1.1](#)) with these parameters and $[\text{Fe}/\text{H}] = -4.0$ (the lowest available abundance in many grids and which should be adequate for the task), a line strength of $20 \text{ m}\text{\AA}$ corresponds to $[\text{Ca}/\text{H}]_{\text{min}} = -9.4$. (For Fe I $3,859.9 \text{ \AA}$, the intrinsically strongest Fe I line in the optical spectrum, a line strength of $20 \text{ m}\text{\AA}$ results in a less stringent limiting abundance of $[\text{Fe}/\text{H}] = -7.2$.) If one were to assume that this hypothetical star had $[\text{Ca}/\text{Fe}] = 0.4$, similar to that found in the most metal-poor stars, its iron abundance would be $[\text{Fe}/\text{H}]_{\text{min}} = -9.8$. This can be taken as a rough estimate of the lowest metallicity practicably detectable.

Even, however, if such a star existed, one should not automatically interpret the above minimum abundance as the value with which it formed, given the possibility of accretion of material from the interstellar medium (ISM) during its $\sim 13 \text{ Gyr}$ lifetime. Using calculations described by [Frebel et al. \(2009\)](#), who compute the amount of material likely to have been accreted onto each of some 470 observed halo main-sequence stars, it was found that during its time on the red giant branch (RGB), the average amount of material accreted onto a star would have increased an initial zero heavy-element abundance to an observed atmospheric value of $[\text{Fe}/\text{H}] = -8.6$, with a dispersion of 0.8 dex. (Here the large dispersion is driven by an extremely strong dependence of the accretion process on the relative velocity of the star with respect to the ISM.) From this information, it follows that a star that formed with $[\text{Fe}/\text{H}]_{\text{min}} = -9.8$ and experienced average ISM accretion would be observed during its RGB evolutionary phase as an object with $[\text{Fe}/\text{H}] \sim -8.6$. Alternatively, given the dispersion in possible accretion histories, one might also say that the probability of finding a star that initially had zero heavy-element abundance (i.e., Population III) and observed today during its RGB phase would have an “accreted” abundance of $[\text{Fe}/\text{H}] = -9.8$ or smaller, is ~ 0.07 .

Having assessed the technical feasibility of finding near-zero-metallicity, low-mass ($M < 1 M_{\odot}$) stars, one needs also to consider potential physical processes that may have played a role in the formation of the most metal-poor stars and which lead to abundances between the current lowest observed level of $[\text{Fe}/\text{H}] \sim -5.5$ and the potentially detectable $[\text{Fe}/\text{H}] = -9.8$. As will be discussed in [Sect. 6.1](#), the critical factor is the cooling mechanisms that determine the contraction and fragmentation of existing gas clouds. Two potentially important cooling mechanisms are noted here, as well as the abundance limits they impose, following [Frebel et al. \(2009\)](#). The first is C II and O I fine-structure line cooling which leads to $[\text{Fe}/\text{H}]_{\text{min}} = -7.3$. The second is the major cooling due to dust grains, for which the limit might be one to two orders of magnitude lower, for example, $[\text{Fe}/\text{H}]_{\text{min}} = -8.0$ to -9.0 . While more detailed knowledge on cooling mechanisms may well change these values, the above discussion shows that one should not be surprised to find stars with metallicities much lower than those of the most-metal-poor stars currently known.

3 Derived Chemical Abundances

3.1 Abundance Determination

3.1.1 One-Dimensional Model Atmosphere Analyses

Most chemical abundance determinations are based on one-dimensional (1D) model stellar atmosphere analyses that assume hydrostatic equilibrium, flux constancy, local thermodynamic equilibrium (LTE), and treat convection in terms of a rudimentary mixing length theory. (In most cases, the configurations are plane parallel, but when necessary spherical symmetry is adopted for giants.) To first order, the basic atmospheric parameters that define the model are effective temperature (T_{eff}), surface gravity ($\log g$), and chemical composition. Given these, one may construct a model atmosphere and compute the emergent flux for comparison with observations. Then, on the assumption that the model well-represents the observed star, when one obtains a good fit between the model emergent flux (in particular the strengths of the atomic and molecular features) and the observed flux, one assumes the chemical abundances of the model correspond to those of the observed star. The student should consult Gray (2005) and Gustafsson et al. (2008) for the concepts associated with the process.


For completeness, it should be briefly noted that T_{eff} and $\log g$ are sometimes derived from atomic and molecular transitions (excitation temperature and ionization equilibrium, respectively) and sometimes from measurements of continuum colors and the strengths of hydrogen Balmer lines and the Balmer Jump. Surface gravity is also often derived from the star's luminosity, T_{eff} , and mass. A basic shortcoming of 1D modeling is that an artificial, second-order, extra line broadening called "microturbulence," over and above the thermal broadening of the models, is always introduced into the formalism to satisfy the requirement that atomic lines of different strength yield the same abundance. This will not be discussed here further, except to say that the need for this "fudge factor" has proved to be unnecessary in the more physically realistic 3D modeling. Finally, best analysis involves an iterative process that demands the adopted T_{eff} , $\log g$, abundances, and microturbulence are consistent with both the adopted model atmosphere and all of the details in the star's spectrum that are sensitive to these parameters.

As discussed in Sect. 1.2.2, the analysis produces stellar atmospheric abundances $\epsilon(X)$ for species X relative to hydrogen, expressed as $\log_{10}\epsilon(X) = \log_{10}(N_X/N_H) + 12$; in most cases, values are published using the bracket notation $[X/H] = \log_{10}(N_X/N_H)_* - \log_{10}(N_X/N_H)_\odot$, which expresses the results relative to solar values. For completeness, it should be noted that the *elemental* abundances derived in this way represent the contribution of all isotopes; additional isotope ratios can only be determined in a few cases (e.g., C). This contrasts nucleosynthesis models, which yield abundances of each individual calculated isotope abundance. (Publicly available model atmospheres and associated atomic and molecular data may be found, e.g., at <http://kurucz.harvard.edu>, <http://vald.astro.univie.ac.at>, and http://www.physics.nist.gov/PhysRefData/ASD/lines_form.html, while codes for the computation of emergent fluxes and determination of chemical abundances may be found, e.g., at <http://www.as.utexas.edu/~chris/moog.html>.) Given the power of modern computers, this is now a mature and straightforward process, and 1D/LTE abundances, $[\text{Fe}/\text{H}]$ and $[\text{X}/\text{Fe}]$, based on high-resolution, high S/N, data are currently available for a large number of metal-poor stars – for example, for resolving power $>20,000$ and $[\text{Fe}/\text{H}] < -2.0$, data exist for some 600 objects. Two comprehensive

compilations of published material are those of Suda et al. (2008) and Frebel (2010), the latter of which will be used in what follows. The precisions of these results are high, typically of order 0.10 dex (26%) and in some cases ~ 0.03 dex (7%).

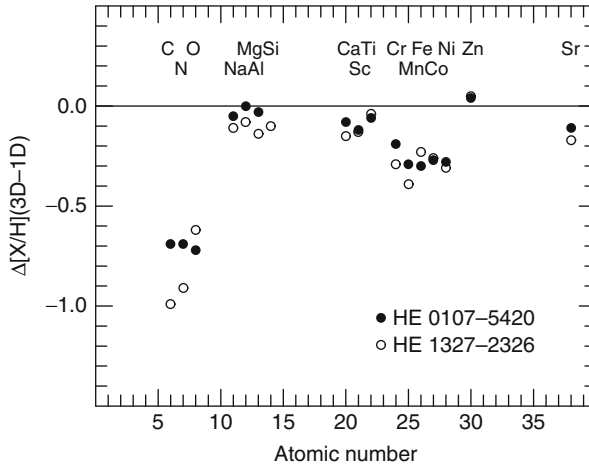
3.1.2 Three-Dimensional Model Atmospheres

The question that remains to be answered is: How accurate are these 1D abundances? The issues have been addressed by Asplund (2005), to whom the reader is referred. Three-dimensional (3D) hydrodynamical models reveal temperature inhomogeneities, which lead to different temperature structures between 3D and 1D models. As metallicity decreases, the inhomogeneities become larger, resulting in significant negative abundance corrections. Figures 1 and 2 of Asplund illustrate the effect: at $T_{\text{eff}} = 5,800$ K, $\log g = 4.4$, one finds that while at $[\text{Fe}/\text{H}] = 0.0$, the *average* temperature of the 3D model agrees reasonably well with that of its 1D counterpart, the situation is very different at $[\text{Fe}/\text{H}] = -3.0$, where the 3D model has temperatures lower by several hundred degrees in its upper layers. These in turn lead to significant differences between 1D and 3D abundances as a function of the metallicity of the star and the excitation potential of the observed line transition. Figure 8 of Asplund shows that for resonance lines of typical elements, 1D abundances are too high by 0.1–0.6 dex, with the difference being smaller for lines of higher excitation potential. The strength of molecular features is very sensitive to the temperatures in the outer layers, resulting in dramatically lower abundances compared with those obtained in 1D analyses.

Abundance analysis utilizing 3D models is a computationally intensive exercise, and results are currently available for only relatively few stars selected for their astrophysical significance. Examples of this are the two most iron-poor stars HE 0107–5240 and HE 1327–2326, for which  Fig. 3-5 presents abundance differences $[\text{X}/\text{Fe}](3\text{D}) - [\text{X}/\text{Fe}](1\text{D})$ versus atomic number from the work of Collet et al. (2006) and Frebel et al. (2008). Note the extremely large differences for C, N, and O – for which the cited results are determined from analysis of CH, NH, and OH, respectively. One must bear this in mind when seeking to interpret 1D chemical abundances.

3.1.3 Departures from Thermodynamic Equilibrium (Non-LTE)

In order to determine chemical abundances, one needs to derive the populations of atomic and molecular energy levels, which depend on details of the radiative and collisional effects in the regions of line formation in the stellar atmosphere. The reader is once again referred to Asplund (2005) for a thorough discussion of this matter. The proper solution to the problem is sufficiently computationally intensive that most investigations to date have made the assumption of LTE. This approach assumes that collisional effects dominate over radiative ones, from which it follows that the required populations can be determined by the Maxwell, Saha, and Boltzmann distributions, which involve only the local physical parameters temperature and electron pressure. To quote Asplund, “In LTE the strength of a line can be straightforwardly predicted from a few properties of the line and the species once the model atmosphere and continuous opacity are known. In non-LTE, in principle everything depends on everything else, everywhere else.” (For completeness, it should also be noted that a remaining uncertainty in current non-LTE analyses is the treatment of inelastic collisions with hydrogen atoms; see Asplund 2005, his Section 2.1.)



■ Fig. 3-5

The difference in abundance, $[X/H]_{3D} - [X/H]_{1D}$, versus atomic number deduced from analyses based on three-dimensional and one-dimensional model atmospheres, for the two most iron-poor stars – the subgiant HE 1327-2326 and the red-giant HE 0107-5420, both with $[Fe/H]_{1D} \sim -5.5$. These stars show the so-far most extreme abundance differences between 1D and 3D analyses (Data from Collet et al. 2006 and Frebel et al. 2008)

Given the time-consuming nature of non-LTE computations, the large majority of abundance analyses to date assume LTE. The advice of Asplund should, however, be recalled. “It is always appropriate to provide the LTE results for comparison purposes, but it is unwise to ignore the available non-LTE calculations when providing the final abundance values.” The present work follows this advice where possible and considers (non-LTE-LTE) differences further in ▶ Sect. 3.2.2.

3.1.4 Caveat Emptor

Two caveats are offered in conclusion. The first is that essentially all of the abundances presented here have been determined using 1D/LTE model atmosphere analyses. In some cases, when 3D and/or non-LTE data are available, comments are included on the resulting differences between the two formalisms. The second point is that for a comprehensive improvement over 1D/LTE results, one needs to use both 3D and non-LTE and not just one of them: in the case of lithium, for example, and as will be discussed in ▶ Sect. 4.1, the 3D and non-LTE corrections are both large, but of opposite sign, and fortuitously largely cancel to give the 1D/LTE result.

3.1.5 Post-Astration Abundance Modification

The final question one must address is whether the abundances obtained from these exhaustive model atmosphere analyses are indeed the values in the protocloud from which the star formed.

Here, very briefly, with source material pertinent to metal-poor stars, are important examples of processes that can modify the original abundance patterns in the observed surface layers:

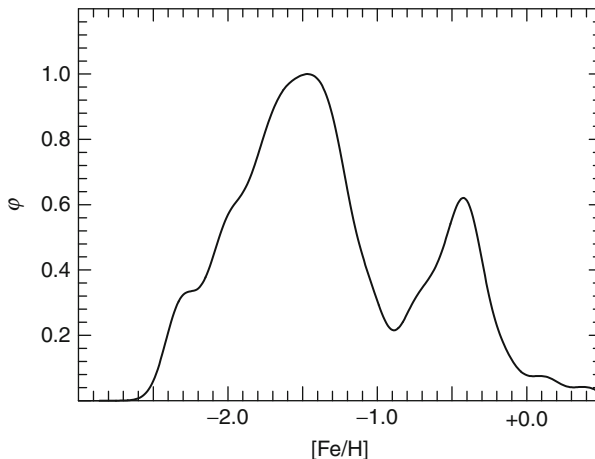
- Accretion from the interstellar medium over the lifetime of a star (e.g., Frebel et al. 2009)
- Radiative and gravitational diffusion in the stellar surface layers (e.g., Behr et al. 1999)
- Macroscopic mixing of nucleosynthesis products from stellar interiors into their surface layers (e.g., Gratton et al. 2000)
- Post-asymptotic-giant-branch evolution, during which the sequence of element fractionation onto circumstellar grains, radiation-pressure-driven grain/gas separation, and the formation of a stellar atmosphere containing the remaining gas produce an Fe-poor, modified abundance pattern determined by the physics of gas/grain condensation (e.g., Giridhar et al. 2005)
- Transfer of material across a multiple stellar system during post-main-sequence evolution (e.g., Beers and Christlieb 2005)

3.2 Abundance Patterns

3.2.1 Metallicity Distribution Functions (MDF)

The Galactic Globular Cluster System

With very few exceptions (which will be considered in [Sect. 4.3.1](#)), the Milky Way's globular clusters are individually chemically homogeneous with respect to iron. The collective MDF of the cluster system is bimodal, as first definitively shown by Zinn (1985) and presented here in [Fig. 3-6](#) (based on the more recent abundance compilation of Carretta et al. 2009).



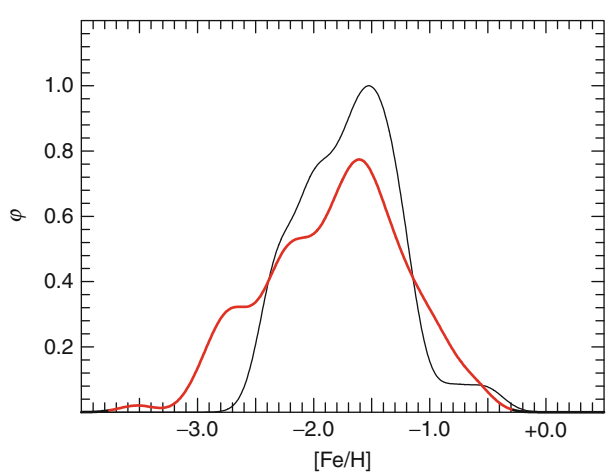
■ Fig. 3-6

MDF of the Galactic globular clusters (data of Carretta et al. 2009). Note the clearly bimodal distribution, with peaks at $[\text{Fe}/\text{H}] = -1.5$ and -0.4 , corresponding predominantly to halo and disk/bulge material, respectively (The histogram was generated with a gaussian kernel having $\sigma = 0.10$ dex)

The two components, initially designated “halo” and “disk,” have mean metallicities $[\text{Fe}/\text{H}] \sim -1.5$ and -0.4 . This terminology, however, appears to be an oversimplification: some clusters with abundances as low as $[\text{Fe}/\text{H}] \sim -1.5$ have disk-like kinematics, some of the inner “disk” subpopulation have been suggested to be members of the Galactic bulge, and consideration of the horizontal branch morphologies of globular clusters first led to the suggestion of old and young subgroups in the halo subpopulation (Zinn 1993). Clearly, the situation is a very complicated one.

Field Stars

MDFs are also available for local metal-poor samples of both kinematically selected main-sequence dwarfs (Carney et al. 1996; Ryan and Norris 1991b), and spectroscopically selected giants (Schörck et al. 2009) and dwarfs (Li et al. 2010).

The field star distributions differ from that of the globular clusters in one important aspect: all halo field star samples contain objects having abundances considerably lower ($[\text{Fe}/\text{H}] = -4.0$ to -3.0) than that of the most metal-poor globular cluster ($[\text{Fe}/\text{H}] \sim -2.5$). According to Carney et al. (1996), the difference between halo clusters and kinematically selected dwarfs is highly significant (at the 93–99.9% level); this effect is shown here in  Fig. 3-7, based on more recent data. For spectroscopically selected samples, on the other hand (which by definition have a strong abundance selection bias toward more metal-poor stars, not present in kinematically selected samples), the significance is less clear. According to Schörck et al. (2009), “A comparison of the MDF of Galactic globular clusters ... shows qualitative agreement with the halo [field star] MDF, derived from the HES, once the selection function of the latter is included. However,

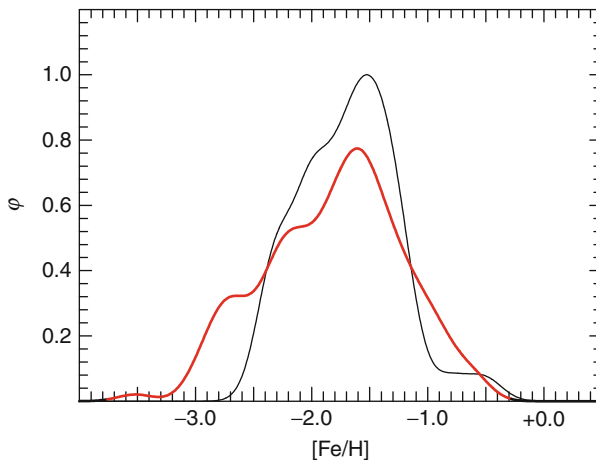
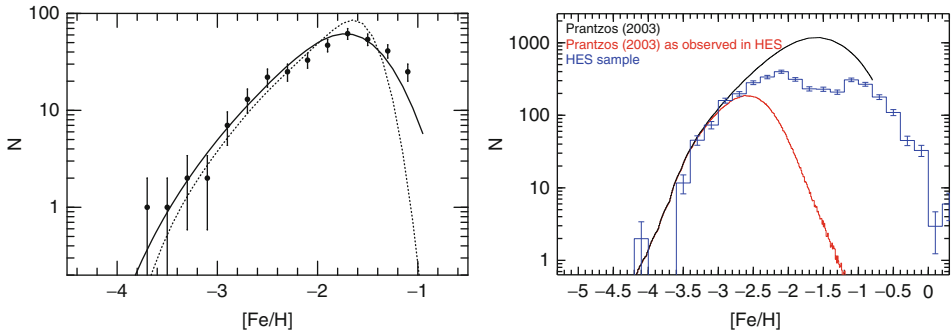


 Fig. 3-7

Comparison of the MDFs of the halo globular clusters (*thin line*) and kinematically chosen halo field main-sequence dwarfs (*thick line*). The selection of the halo samples follows Sect. 3.5 of Carney et al. (1996): for the clusters, only objects more than 8 kpc from the galactic center are included (with abundances from Carretta et al. 2009), while for the dwarfs, the data are from Carney et al. (1996) (The histograms were generated with a gaussian kernel having $\sigma = 0.15$ dex)



■ Fig. 3-8

Left: Comparison of the MDF of the kinematically selected halo main-sequence dwarf sample of Ryan and Norris (1991b) with the simple model (dotted line) of Hartwick (1976) and the more realistic supernova-induced star-formation model (solid line) of Tsujimoto et al. (1999). (The figure was taken from Tsujimoto et al. 1999.) **Right:** MDF of the spectroscopically chosen halo giant sample of Schörck et al. (2009) in comparison with the GCE model of Prantzos (2003). Note the large correction necessary to modify the model (the upper continuous line) for the abundance incompleteness caused by the spectroscopic selection function (the lower continuous line) (from Schörck et al. 2009)

statistical tests show that the differences between these are still highly significant.” The problem with the spectroscopically chosen HK and HES samples is that the corrections that must be applied to compensate for the selection function are very large, as clearly shown in [Fig. 3-8](#) (see also Schörck et al. 2009, their Fig. 17).

The fundamental importance of metallicity distribution functions is that they provide essential constraints on galactic chemical enrichment (GCE) models. The starting point adopted for the present discussion is the simple model of halo chemical enrichment of Hartwick (1976), who assumed that initially, the halo contained zero heavy elements and was chemically enriched by the ejecta of massive stars on timescales short compared with those of the halo’s dynamical evolution (instantaneous recycling). He also assumed that the initial mass function was constant with time and in order to reproduce the MDF of the halo globular clusters, postulated that gas was removed from the system at a rate proportional to that of star formation. The left panel of [Fig. 3-8](#) shows a comparison of this simple model (dotted line) with the observations of halo field dwarfs by Ryan and Norris (1991b). The solid line in the figure, which somewhat better fits the data, represents a model of Tsujimoto et al. (1999) which involves star formation on shells swept up by the ejecta of the supernova explosions of massive stars.

A point worth reiterating from [Sect. 2.2](#) is that the simple Hartwick model predicts the number of metal-poor stars should decrease by a factor of 10 for each factor of 10 decrease in abundance: Norris (1999) and Schörck et al. (2009) report that this appears to be the case down to $[Fe/H] = -4.0$ and -3.6 , respectively, below which there is a large dearth of stars. Recall also from [Sect. 2.4](#) that only four stars with $[Fe/H] \lesssim -4.3$ are currently known.

Several other GCE models have been proposed which modify the basic assumptions of the Hartwick model. As an example, the right panel of [Fig. 3-8](#) shows the comparison between the spectroscopically selected halo giant sample of Schörck et al. (2009) and the model of Prantzos (2003) (which investigates improvement of the instantaneous recycling approximation

and possible gaseous infall). The reader is referred to Schörck et al. (2009) and Li et al. (2010) for comparison of the observations with other GCE models from T. Karlsson (delayed chemical enrichment at the earliest times), S. Salvadori and coworkers (Λ CDM framework with a critical metallicity for low-mass star formation), and N. Prantzos (semi-analytical model within the hierarchical merging paradigm).

This section is concluded with a caveat concerning the above comparison of MDFs. There has been growing evidence over some two decades, beginning with the seminal works of Hartwick (1987) and Zinn (1993) that the Galactic halo comprises more than one component, with different properties as a function of Galactocentric distance; see Carollo et al. (2010) and Morrison et al. (2009) and references therein for details. The multiplicity of the Galactic halo will be discussed in [◆ Sect. 6.2](#). Suffice it here to say it makes little sense to compare the MDFs of samples (observational and/or theoretical) that have different properties (except to test the null hypothesis). It is essential to match the underlying characteristics of the theoretical models and observed samples that are being compared.

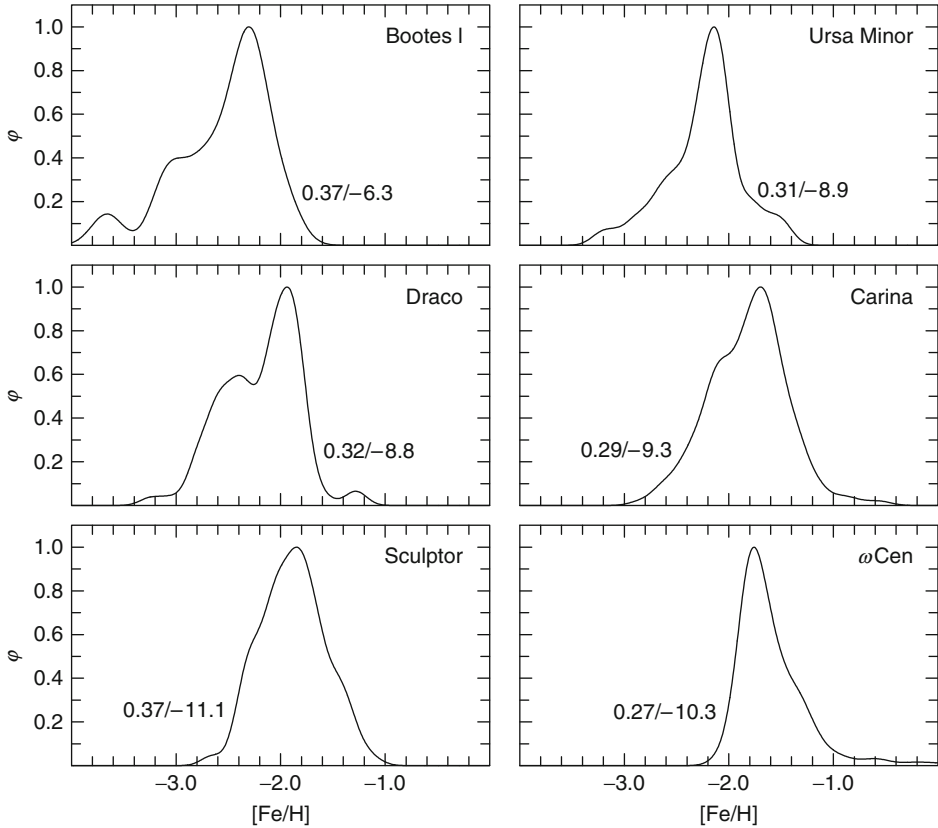
Dwarf Spheroidal Galaxies (dSph)

In stark contradistinction to the Milky Way's globular clusters, its dwarf spheroidal galaxy satellites all show large internal spreads in the abundance of iron. The MDFs for 5 of the ~ 25 currently known systems are shown in [◆ Fig. 3-9](#). (Also shown in the figure, for comparison purposes, is the MDF of ω Cen, the most massive Milky Way globular cluster, and one of only ~ 5 halo clusters known to exhibit a dispersion in iron greater than $\sigma[\text{Fe}/\text{H}] \sim 0.03$ dex.) In each panel, the abundance dispersion $\sigma[\text{Fe}/\text{H}]$ and integrated absolute visual magnitude, $M_{V, \text{total}}$, are also shown.

It has been known for some time that the metallicities of elliptical galaxies and the more luminous dSphs collectively decrease as luminosity decreases (Mateo 1998). As reported by Kirby et al. (2008), the mean $[\text{Fe}/\text{H}]$ of dSphs continues to decrease with decreasing luminosity over the range of the ultra-faint systems as well. As shown here in [◆ Fig. 3-10](#), the relationship holds over the range $3.5 \lesssim \log(L_{\text{tot}}/L_{\odot}) \lesssim 7.5$. This is a clear signal that the dwarf galaxies have undergone internal chemical evolution. Examination of [◆ Fig. 3-9](#) also shows that in the faintest of the dwarf systems ($M_{V, \text{total}} \lesssim -7$), there is a large fraction of stars with $[\text{Fe}/\text{H}] < -3.0$, suggesting a relationship between ultra-faint dwarf galaxies and the most metal-poor stars in the Milky Way halo. This topic will be further addressed in [◆ Sect. 4.3.2](#), but it is worth noting here that an essential difference between the Milky Way's globular clusters and dSph systems is that (for objects with integrated magnitudes $M_{V, \text{total}} \gtrsim -10$) the dSphs are embedded in dark-matter halos (with $M/L_V \sim 10\text{--}10^4$ in solar units), while almost all clusters contain relatively little or no dark matter ($M/L_V \lesssim 5$). This is almost certainly the essential difference behind the large $[\text{Fe}/\text{H}]$ dispersions observed in the dSph systems but absent from the globular clusters.

3.2.2 Relative Abundances

Just as $[\text{Fe}/\text{H}]$ is adopted as proxy for a star's overall metallicity, the abundances of the other elements are most often expressed relative to Fe, i.e., as $[X/\text{Fe}]$ for element X. (This is a somewhat arbitrary definition, driven by the practicality of the richness of the Fe I spectrum, and from time-to-time, the implications of adopting an alternative element as reference are investigated. For an example of this, see Cayrel et al. (2004).) Element abundances are thus directly related to the element that represents the end stage of stellar evolution and provides a good indicator

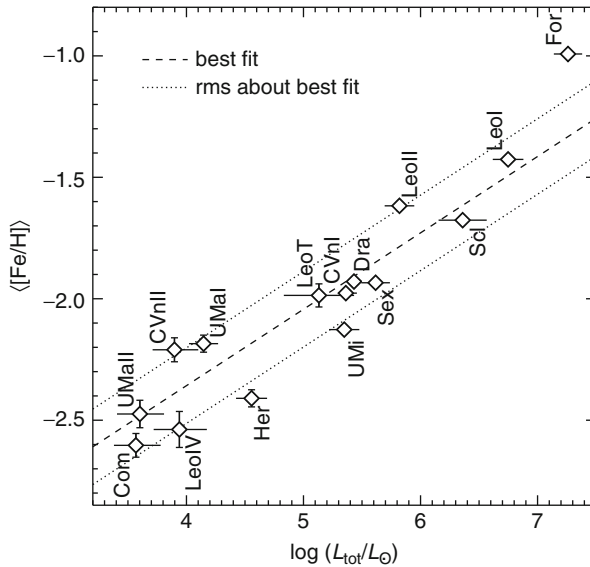


■ Fig. 3-9

The metallicity distribution functions for five Milky Way dwarf galaxies and the globular cluster ω Centauri. Also shown in each panel are $\sigma[\text{Fe}/\text{H}]/M_{v, \text{total}}$ (the dispersion in $[\text{Fe}/\text{H}]$ and the integrated absolute visual magnitude of the system). See Norris et al. (2010b) for source material (The histograms were generated with gaussian kernels having $\sigma = 0.10 - 0.15$ dex)

of core-collapse SN nucleosynthesis. Note that all such (relative) abundances are relative not only to Fe, but also to the abundances measured for the Sun (the bracket notation). This should be kept in mind when “reading” the chemical relative abundance trends in metal-poor stars in terms of galactic chemical evolution.

To give the reader a feeling for the scope of the observed trends, ● Fig. 3-11 shows the 1D/LTE relative abundances for metal-poor Galactic halo red giants from the work of Cayrel et al. (2004), Spite et al. (2005), and François et al. (2007), which covers the range $-4.5 < [\text{Fe}/\text{H}] < -2.0$ and is regarded by many as the “gold standard” of the state of the art for this type of work. The reader should note that the scale in 16 of the 18 panels of ● Fig. 3-11 is the same, with a range in $[\text{X}/\text{Fe}]$ of 2 dex. For the remaining two cases ($[\text{Sr}/\text{Fe}]$ and $[\text{Ba}/\text{Fe}]$), this is insufficient to cover the range in the early Universe, and for these, the relevant panel range is 5 dex! The dotted lines in the figure correspond to the solar value. The solid lines in the



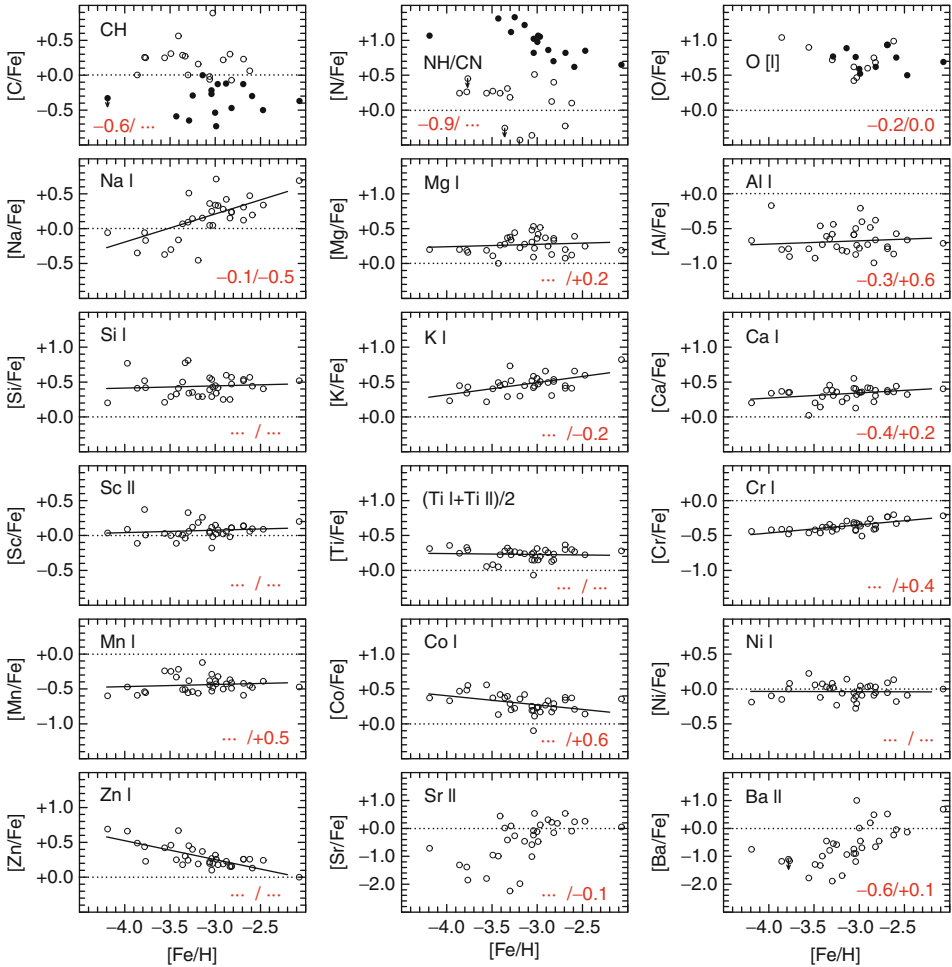
■ Fig. 3-10

Mean $[Fe/H]$ vs. luminosity for the Milky Way's dwarf spheroidal galaxies. Systems with $L_V \leq 10^5 L_{\odot}$ are designated “ultra-faint” dwarf galaxies, since they are fainter than the long-known “classical” dwarf galaxies. It is very likely, however, that there exists a continuous transition between the physical properties of the two groups (Prepared by E. N. Kirby using data from Kirby et al. 2008, and references therein)

panels of sodium through zinc represent the regression lines of Cayrel et al. (2004); for these elements, these authors report errors of measurement $\sigma \sim 0.05\text{--}0.10$ dex and dispersions about the regressions of $1 - 2\sigma$. Also shown, at the bottom right of each panel, are indicative 3D/non-LTE corrections for stars with $[Fe/H] \sim -3.0$ that have been gleaned from Asplund (2005) and other sources in the literature (such as the works of S. M. Andrievsky, D. Baumüller, M. Bergemann, D. V. Ivanova, L. Mashonkina, and co-workers).

By way of introduction to what follows, several aspects of the figure are highlighted:

- The large spreads in C, N, Sr, and Ba are real and tell us much about internal mixing during red giant evolution (C and N: [▶ Sect. 4.2.1](#)) and the processes that produce the heavy neutron-capture elements (Sr and Ba: [▶ Sect. 4.2.2](#)).
- Systematic enhancements of the α -elements Mg, Si, Ca, and (partially) Ti lead to an explanation involving SNe of type Ia and II, operating at different times ([▶ Sects. 4.2.1](#) and [▶ 4.3.2](#)).
- Tight solar-like correlations, such as those of the iron-peak elements Sc and Ni, suggest a close relationship between the production mechanisms of some of the iron-peak elements, indicative of processes similar to those responsible for the enrichment of the Sun.
- In contrast to the previous point, the trends shown by iron-peak elements such as Cr, Mn, and Co cast doubt on the previous suggestion. This “contradiction” is indicative perhaps of



■ Fig. 3-11

1D/LTE relative abundances ($[X/Fe]$) versus $[Fe/H]$ for metal-poor halo red giants from the work of Cayrel et al. (2004), Spite et al. (2005), and François et al. (2007). In the *top row*, filled and open circles refer to “mixed” and “unmixed” stars, respectively, as defined by Spite et al. (see Sect. 4.2.1). Also shown at the *bottom* of each panel are indicative (3D–1D)/(non-LTE–LTE), abundance differences as discussed in Sect. 3.1.2 (“...” indicates a potential incompleteness in our literature search or the absence of relevant information)

- differences related to the location of the mass-cut radius within the progenitor of the SN explosion (above which all material is expelled) or to non-LTE effects (Sect. 4.2.1).
- Large corrections to some of the 1D/LTE abundances are clearly necessary to take into account 3D effects and a more realistic treatment of non-LTE before they may be closely and reliably compared with the prediction of stellar evolution and GCE computations.

4 The Chemical Evolution of the Universe

4.1 Relics of the Big Bang


According to Standard Big Bang Nucleosynthesis (SBBN), some minutes after the singularity at the era of decoupling of radiation and matter, the only chemical elements in the Universe were hydrogen, helium, and lithium. With the additional constraint of the results of the Wilkinson Microwave Anisotropy Probe (WMAP), the predicted relative mass densities of these elements are 0.75, 0.25, 2.3×10^{-9} , respectively (Spergel et al. 2007). All other elements have been produced subsequently.


4.1.1 Helium

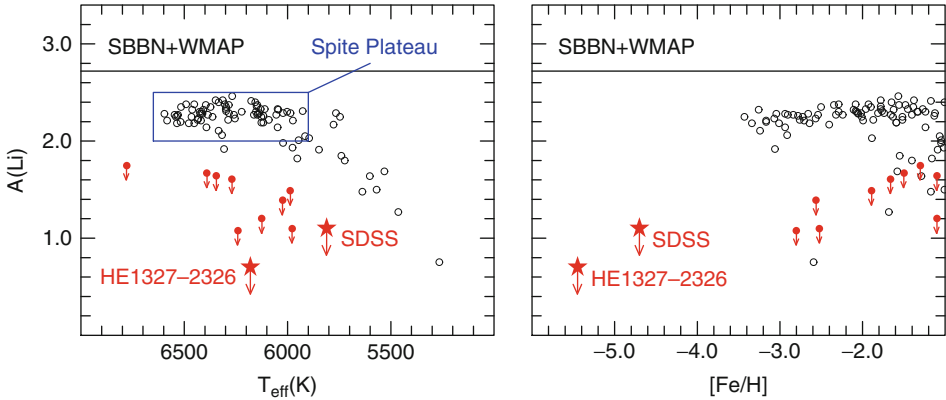
No reliable spectroscopic determinations exist of the abundance of helium in the atmospheres of stars having $[\text{Fe}/\text{H}] < -1.0$. Most are too cool ($T_{\text{eff}} < 7,000$ K) for lines of neutral helium to be currently useful for abundance analysis (see, e.g., Dupree et al. 2011), and in metal-poor stars hot enough for the test to be made (the so-called blue-horizontal-branch stars), strong diffusive processes are clearly at work in their outer layers and preclude determination of the chemical abundances in the material from which they formed (Behr et al. 1999). The best estimates of primordial helium abundance based on spectral features of helium come from the analysis of helium lines in gaseous nebulae, for which Spergel et al. (2007) report a primordial helium abundance $Y_p = 0.232 - 0.258$.

4.1.2 Lithium

Spite and Spite (1982) first demonstrated that the Li abundance of metal-poor, near main-sequence-turnoff stars appears constant in the temperature range $T_{\text{eff}} = 5,500\text{--}6,250$ K, and concluded, “the abundance of lithium at the beginning of the Galaxy was: $N_{\text{Li}} = 11.2 (\pm 3.8) 10^{-11} N_{\text{H}}$,” i.e., $A(\text{Li}) = 2.05 \pm 0.16$. Astronomers today discuss this fundamental discovery not so much in respect of the Galaxy, to which it is certainly pertinent, but rather in terms of the Li abundance that emerged from the Big Bang.

The effect is shown in  *Fig. 3-12*, based on more recent observational data, where 1D/LTE values of $A(\text{Li})$ are presented as a function of T_{eff} and $[\text{Fe}/\text{H}]$. (It is noted here for completeness that the accepted temperature scale for metal-poor main-sequence stars has become some 300 K hotter since the work of Spite and Spite (1982).) One sees that the so-called Spite plateau remains clearly defined, and as appreciated by Spite and Spite (1982), for $T_{\text{eff}} < 5,900$ K on the new scale, lithium is destroyed by strong convective circulation that brings it into deeper and hotter regions. During extensive expansions on the original sample, some stars were found in which Li could not be detected. It has been suggested that these stars, which comprise only a small fraction of their parent population, have ultra low lithium abundances as the result of phenomena related to binarity and blue stragglers, during which Li is converted into other elements at the high temperatures experienced during convective mixing in their outer layers (see Ryan et al. 2001).

There are, however, two extremely important further points to be taken from  *Fig. 3-12*. The first is that, ignoring the obvious outliers, the mean abundance of the plateau,



■ Fig. 3-12

1D/LTE lithium abundance, $A(\text{Li})$, as a function of T_{eff} (left) and $[\text{Fe}/\text{H}]$ (right), from the work of Frebel et al. (2008, HE 1327–2326) and Caffau et al. (2011, SDSS J102915+172927) (large filled stars), Meléndez et al. (2010, open circles), and Ryan et al. (2001, filled circles). The horizontal line in each panel is the value predicted from the observations by WMAP interpreted in terms of SBBN (Cyburt et al. 2008)

$A(\text{Li}) = 2.28 \pm 0.01$, lies some 0.4–0.5 dex below the value that has been predicted from the results of WMAP, interpreted in terms of the predictions of SBBN of $A(\text{Li}) = 2.72^{+0.05}_{-0.06}$ (Cyburt et al. 2008). (Here the error in the mean abundance of the plateau admits no slope or step as a function of T_{eff} , both of which have been claimed in the literature.)

The second point is that for the most metal-poor near-main-sequence-turnoff stars, HE 1327–2326 and SDSS J102915+172927 with $[\text{Fe}/\text{H}]_{\text{ID,LTE}} = -5.4$ and -4.7 , lithium is not detected, leading to the extremely puzzling limits of $A(\text{Li}) < 0.7$ and < 1.1 , respectively. Given that these objects have $T_{\text{eff}} = 6,180$ and $5,810$ K, one would have expected them to lie on the Spite plateau. (Note also that there is no evidence yet for binarity, or any other (non-abundance) peculiarity, for these stars.) This question will be addressed further in ▶ Sect. 6.1, where they are discussed in more detail.

Before proceeding, it should be noted that available 3D/non-LTE computations appear to be in agreement with those based on the 1D/LTE assumptions. The reader is referred to Asplund et al. (2003), who report that for two stars (with $T_{\text{eff}}/\log g/[\text{Fe}/\text{H}] = 5,690/1.67/-2.50$ and $6,330/2.04/-2.25$), the 3D and non-LTE corrections are both large, with absolute values of ~ 0.3 dex, but of opposite sign, which essentially cancel to yield a total correction of only ~ 0.05 dex. That is to say, 1D/LTE Li abundances are fortuitously valid.

Given the accuracy of the WMAP/SBBN prediction of the primordial Li abundance, the most widely held view seems to be that the abundance obtained from the analysis of observed Li line strengths in near-main-sequence-turnoff metal-poor stars is not the primordial value and that an explanation of the difference will lead to a deeper understanding of the astrophysics of stars and galaxies. The reader should consult Korn et al. (2007), Lind et al. (2009), and Meléndez et al. (2010) for recent examples of this approach, based on lithium abundances of

field (Meléndez et al.) and globular cluster (Korn et al. and Lind et al.) near-main-sequence-turnoff stars. Meléndez et al. (2010) report, “Models including atomic diffusion and turbulent mixing seem to reproduce the observed Li depletion ... which agrees well with current predictions from ... standard Big Bang nucleosynthesis,” while Lind et al. (2009) state “We confirm previous findings that some turbulence, with strict limits to its efficiency, is necessary for explaining the observations.” Both, on the other hand, issue a *caveat emptor*: “We caution however that although encouraging, our results should not be viewed as proof of the ... models until the free parameters required for the stellar modeling are better understood from physical principles” (Meléndez et al. 2010) and “However, these models fail to reproduce the behavior of Li abundances along the plateau, suggesting that a detailed understanding of the physics responsible for depletion is still lacking” (Lind et al. 2009).

4.2 The Milky Way Halo

The evolution of the chemical elements began shortly after the Big Bang and is an ongoing process. It can be traced in detail in the Milky Way with stars of different metallicities, ranging from the most metal-deficient to the most metal-rich. Iron abundance serves as proxy not only for the overall metallicity of a star but also for the evolutionary timescales it took to enrich the gas from which stars formed. It is not possible in most cases, however, to determine the ages of individual field stars, and what is known of Milky Way halo ages is derived from the fitting of globular cluster and field star near-main-sequence-turnoff color-magnitude diagrams to stellar evolution modeling and nucleo-chronometry of metal-poor field stars. (This topic is further addressed in [♦ Sect. 5](#).) As noted in [♦ Sect. 1.3](#), the present discussion is restricted principally to stars of the Galactic halo having $[\text{Fe}/\text{H}] < -1.0$. The Galactic age-metallicity relationship suggests that it took of order ~ 4 Gyr to reach this abundance (see e.g., Freeman and Bland-Hawthorn 2002). For comparison, the (one zone) galactic chemical enrichment model of Kobayashi et al. (2006) also takes ~ 4 Gyr to reach $[\text{Fe}/\text{H}] = -1.0$. The abundance trends discussed in the following thus describe the first ~ 5 Gyr of the evolution of the Milky Way – which, for the present discussion, is taken as a first approximation to the timescale for a similar enrichment of the Universe.

To understand the production of the elements and the observed trends found for metal-poor stars as a function of overall metallicity, most subsections below begin with a description of the relevant nucleosynthesis processes. Arnett (1996) and Wallerstein et al. (1997) provide general introductions to this topic. Woosley and Weaver (1995), among others, have carried out extensive core-collapse SN yield calculations to investigate the synthesis of the different isotopes during stellar evolution and subsequent supernova explosion. Progenitor masses of 11–40 M_{\odot} and different metallicities were considered. Since the details of the explosion mechanism of SNe remain largely unknown, a piston approximation (for the sudden injection of energy – the “explosion”) is employed so that the post-SN nucleosynthesis can be calculated. Fortunately, relatively few isotopes appear to be significantly affected by this uncertainty. The overall explosion energy and the “mass cut” (a specific radius above which material is ejected, rather than falling back onto the nascent black hole or neutron star) thus have significant impact on the final abundance distribution.


Mainly intermediate-mass elements, with $Z \leq 30$, are produced and ejected by core-collapse supernovae. Traces of the so-called neutron-capture elements ($Z \gtrsim 30$) are believed to be produced by SNe and also during the asymptotic-giant-branch (AGB) phase of evolution of


low- and intermediate-mass ($\sim 1\text{--}8 M_{\odot}$) stars. These heavier elements are about one million times less abundant than the lighter ones. Irrespective of their quantities, however, all elements play an important role in our understanding of galactic chemical evolution since each reflects the interplay of all the astrophysical processes and sites that produced the elements as they are known today.


In what follows, an unfortunately somewhat incomplete discussion is presented of many of the elements that are observed in metal-poor stars. Reasons for the incompleteness range from simple space limitations of this chapter to the fact that not all chemical elements can be observed in the relatively cool main-sequence and giant stars reviewed here.

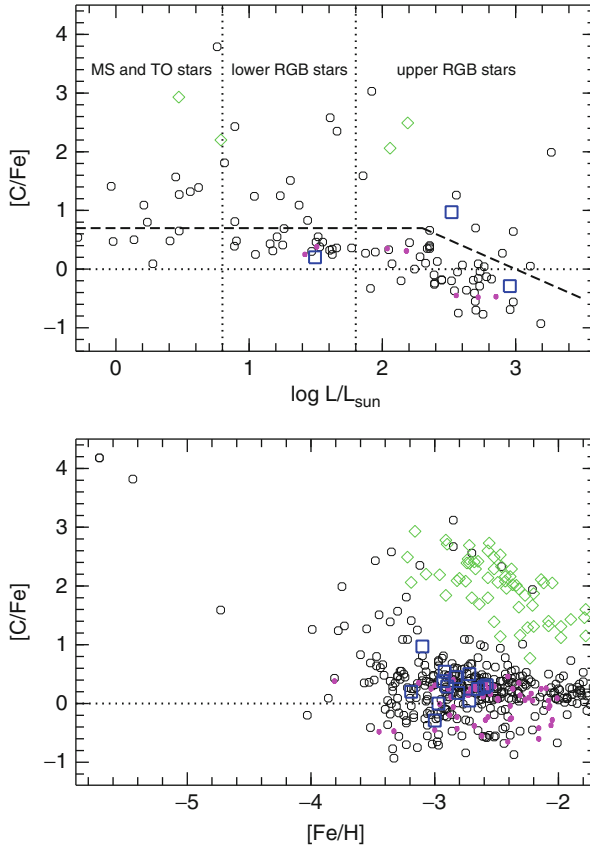
4.2.1 The Evolution of Carbon Through Zinc

Carbon, Nitrogen, and Oxygen

C, N, and O are synthesized during quiescent stellar evolution and in Type II (core-collapse) SN explosions. Carbon is produced in the triple- α process during advanced evolutionary stages such as the AGB phase. It is released into the interstellar medium during supernova explosions if the star is massive enough, or through stellar winds, if significant. Whenever the triple- α process is at work, some oxygen is created as a by-product in the α -process. Hence, O can be regarded as an “ α -element” (see below), and abundance studies have shown that O does indeed exhibit this behavior in metal-poor stars (see  Fig. 3-11, top right, and the discussion below). Nitrogen is produced during H-burning in the CNO-cycle, during which C, N, and O act as catalysts. In this process, C and O abundances decrease while N increases, as the CNO-cycle approaches equilibrium. At the same time, the $^{12}\text{C}/^{13}\text{C}$ ratio is driven to small values (~ 4). The production of nitrogen can be increased by stellar rotation: fast rotating, massive Population III stars (Meynet et al. 2006) may, for example, have been significant producers of the first enrichments in CNO elements.

In metal-poor stars, the abundances of each of C, N, and O can be determined from observations of their hydrides – the G-band of CH at $\sim 4,300 \text{ \AA}$, the near-UV NH feature at $3,360 \text{ \AA}$, and the UV features of OH at $3,100 \text{ \AA}$. CN and/or C_2 bands can also provide constraints at optical wavelengths. The point that must be repeated here is that (as noted in  Sect. 3.2.2) 1D/LTE abundances determined from CH and NH may overestimate C and N abundances by up to ~ 0.7 dex. Atomic features of CI at $\sim 9,070 \text{ \AA}$, together with the forbidden O I line at $6,300 \text{ \AA}$ and the O I triplet at $\sim 7,770 \text{ \AA}$, provide other important constraints on the abundances of these elements. The three diagnostics involving oxygen yield different 1D/LTE abundances, driven by 3D and non-LTE effects. Again, the reader should consult Asplund (2005) for a thorough discussion of the problem: suffice it here to say that only for the forbidden O I line are the 1D/LTE abundances relatively unaffected. Recent results, taking into account various abundance corrections (e.g., Fabbian et al. 2009), indicate relatively small variation of [O/Fe] as a function of [Fe/H] in metal-poor stars.

Evolutionary mixing effects also modify initial surface abundances. Dredge-up events and mixing bring nuclei from interior layers to the surface, including CNO-processed material. The surface abundances of heavier elements are not affected by these mixing processes, and their relative fractions remain unchanged. The effect is shown in  Fig. 3-11 where one sees a clear anti-correlation between C and N in “mixed” stars (filled circles; [C/Fe] < 0.0 and [N/Fe] > +0.5) and “unmixed” stars (open circles; [C/Fe] \geq 0.0 and [N/Fe] < +0.5), as defined by Spite et al. (2005).

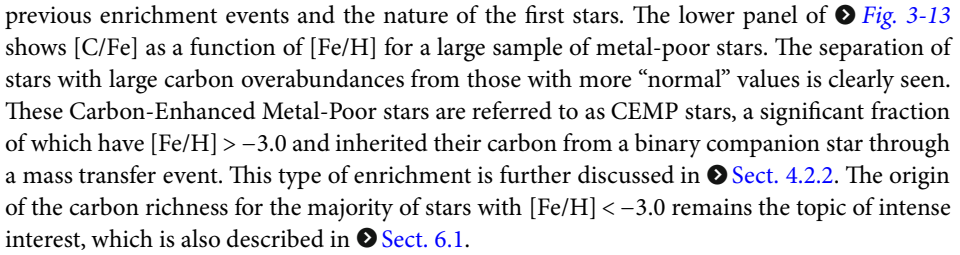
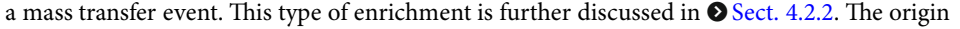
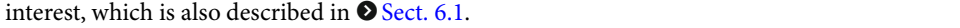


■ Fig. 3-13

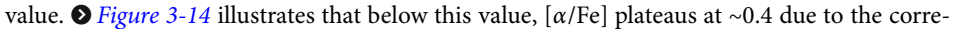
1D/LTE $[C/Fe]$ abundances as a function of $[Fe/H]$ and luminosity (data from the Frebel 2010 compilation). In the *top panel*, only stars with $[Fe/H] < -3.0$ are included. The definition for C-rich objects of $[C/Fe] > 0.7$ but with a luminosity dependent decline reflecting internal mixing processes is shown as a *dashed line* (see also [Table 3-1](#)). *Open diamonds* are used for s- and r+s-process-rich metal-poor stars, *open squares* refers to r-II and small *filled circles* to r-I r-process-rich objects, which are further discussed in [Sect. 4.2.2](#)


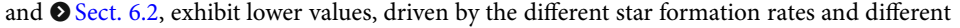
These mixing effects are observed in metal-poor stars with increasing luminosity on the upper RGB, as illustrated in the upper panel of [Fig. 3-13](#), where only stars with $[Fe/H] < -3.0$ are shown. A downturn of the $[C/Fe]$ ratio can be clearly seen at luminosity $\log L/L_{\odot} > 2$. When studying the CNO group, stellar evolutionary status must therefore be taken into account.

Many metal-poor stars show C abundances well in excess of the general trend set by most Population II stars (and unrelated to abundance changes due to mixing) at all metallicities. Such “extra” carbon must have come either from additional sources that enriched the material from which the star formed or from enriched material that was added to the star at a later time. Carbon measurements in metal-poor stars thus provide important information on the various

previous enrichment events and the nature of the first stars. The lower panel of  Fig. 3-13 shows $[C/Fe]$ as a function of $[Fe/H]$ for a large sample of metal-poor stars. The separation of stars with large carbon overabundances from those with more “normal” values is clearly seen. These Carbon-Enhanced Metal-Poor stars are referred to as CEMP stars, a significant fraction of which have $[Fe/H] > -3.0$ and inherited their carbon from a binary companion star through a mass transfer event. This type of enrichment is further discussed in  Sect. 4.2.2. The origin of the carbon richness for the majority of stars with $[Fe/H] < -3.0$ remains the topic of intense interest, which is also described in  Sect. 6.1.

α -Elements

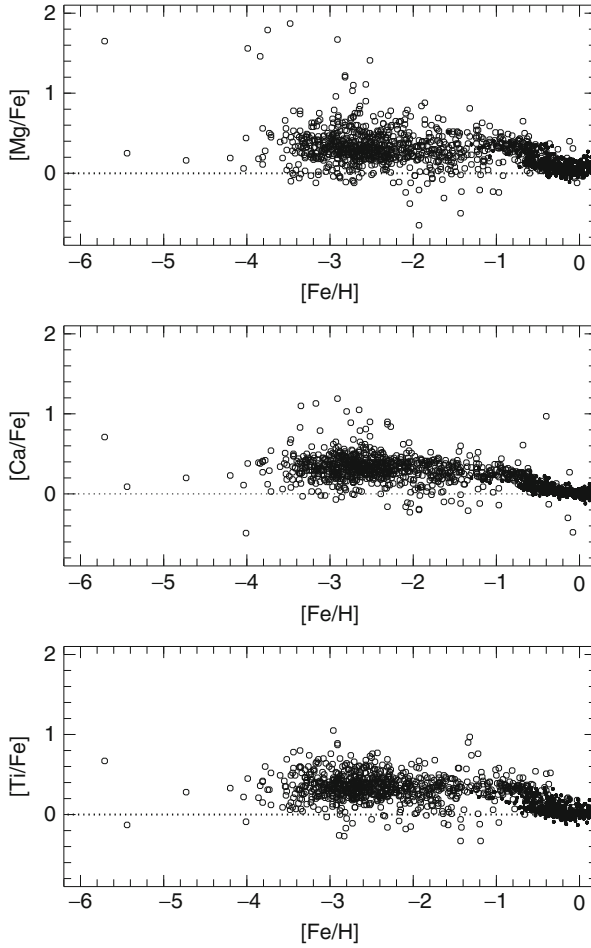
The α -elements (Mg, Ca, Si, Ti) are built from multiples of He nuclei via α -capture during various stages of stellar evolution (carbon burning, neon burning, complete, and incomplete Si burning). Although Ti ($Z = 22$) is not a true α -element, in metal-poor stars, the dominant isotope is ^{48}Ti , which behaves like one. Produced in massive stars, these α -element nuclei are then dispersed during subsequent SN explosions. Abundance studies have shown that the majority of metal-poor stars with $[Fe/H] < -1.0$ shows an enhanced $[\alpha/Fe]$ ratio compared with the solar value.  Figure 3-14 illustrates that below this value, $[\alpha/Fe]$ plateaus at ~ 0.4 due to the correlated production and release of α -elements and Fe. This characteristic overabundance in halo stars reflects an enrichment by core-collapse SNe in the early Universe. At later times (roughly 1 Gyr after the Big Bang), once the first lower-mass stars reached the end of their lifetimes, SN Ia explosions began to dominate the production of Fe. The main yield of SNe Ia is C-, O-, and Fe-peak elements. This change in Fe producers can be seen in the abundance trends of metal-poor stars. Above metallicities of $[Fe/H] \sim -1.0$, the onset of SNe Ia and their Fe contribution to the chemical evolution of the Milky Way manifests itself in a pronounced decrease of the stellar $[\alpha/Fe]$ values (e.g., Ryan et al. 1996) until $[\alpha/Fe] = 0.0$ is reached at $[Fe/H] = 0.0$.

There are important exceptions to this generalization. Some metal-poor stars show large Mg and Si abundances possibly due to unusual supernova explosions and associated nucleosynthesis processes (Aoki et al. 2002; Frebel et al. 2005). Others, as will be discussed in  Sect. 4.3.2 and  Sect. 6.2, exhibit lower values, driven by the different star formation rates and different relative α/Fe contributions from Type II and Type Ia supernovae.

The α -elements also serve to highlight a further potentially important role of relative abundances as a function of $[Fe/H]$. Because the abundance of a given element contains the history of all the SNe that have contributed to the cloud from which a star forms, the dispersion of observed relative abundances contains potentially strong constraints on the relative yields of SNe, the stellar mass function, and the efficiency with which the ejecta of SNe have been mixed with the existing ISM. Several authors have emphasized the small values of the dispersion in $[Mg/Fe]$ (~ 0.06 – 0.10 dex) in homogeneously selected and analyzed halo samples, which lead to interesting restrictions on the above possibilities.

Iron-Peak Elements

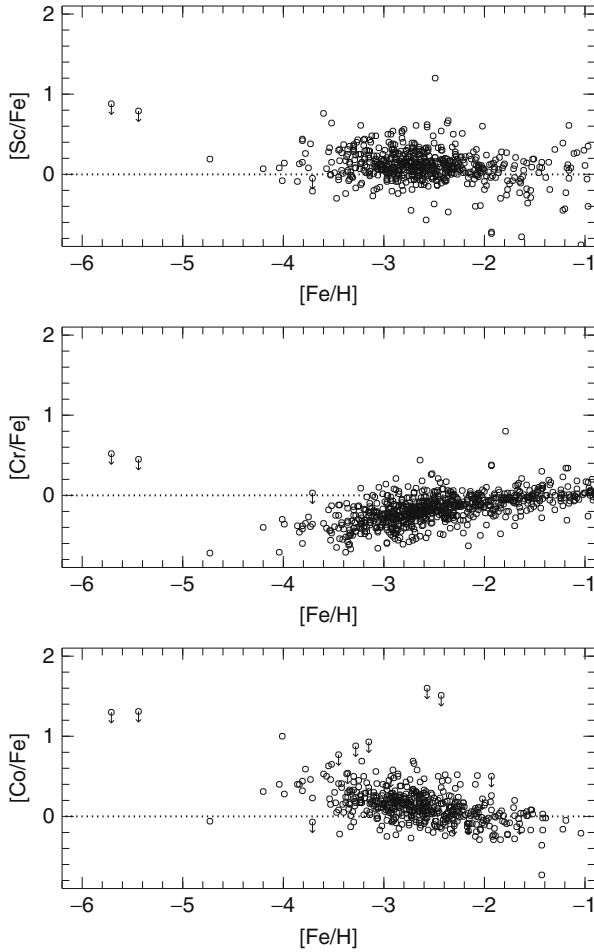
In the early Universe, the iron-peak elements (V to Zn; $23 \leq Z \leq 30$) were synthesized during the final evolution of massive stars, in a host of different nucleosynthetic processes before and during core-collapse SN explosions. These include direct synthesis in explosive burning stages (explosive oxygen and neon burning, and complete and incomplete explosive Si burning), radioactive decay of heavier nuclei or neutron-capture onto lower-mass Fe-peak elements during helium and later burning stages, and α -rich freeze-out processes.



■ Fig. 3-14

1D/LTE α -element ratios $[\text{Mg}/\text{Fe}]$, $[\text{Ca}/\text{Fe}]$, and $[\text{Ti}/\text{Fe}]$ as a function of $[\text{Fe}/\text{H}]$. *Open circles* denote halo stars from the compilation of Frebel (2010). *Small dots* are more metal-rich thin and thick disk stars from the Venn et al. (2004) compilation. At $[\text{Fe}/\text{H}] > -1.0$, all three element ratios decrease with respect to the average halo value of $[\alpha/\text{Fe}] \sim 0.4$. This is due to the onset of SNe Ia, which contribute relatively large amounts of Fe to the chemical enrichment of the Galaxy. By $[\text{Fe}/\text{H}] \sim 0.0$, the solar ratio of $[\alpha/\text{Fe}]$ is reached

In ► Fig. 3-15, 1D/LTE relative abundances, $[\text{Sc}/\text{Fe}]$, $[\text{Cr}/\text{Fe}]$, and $[\text{Co}/\text{Fe}]$, are presented as a function of $[\text{Fe}/\text{H}]$, which demonstrate quite different behavior as $[\text{Fe}/\text{H}]$ increases. ► Fig. 3-11 also presents data for other Fe-peak elements (Mn, Ni, and Zn) on the range $-4.0 \lesssim [\text{Fe}/\text{H}] \lesssim -2.5$. The abundance trends of Cr and Mn have a pronounced positive slope: their abundances at the lowest metallicities are subsolar ($[\text{Cr}, \text{Mn}/\text{Fe}] \sim -0.5$ at $[\text{Fe}/\text{H}] \sim -3.5$), becoming solar-like at $[\text{Fe}/\text{H}] \sim -1.0$. In contradistinction, the Co and Zn abundance trends are in the opposite sense. Their abundances decrease from $[\text{Co}, \text{Zn}/\text{Fe}] \sim +0.5$ at



■ Fig. 3-15

1D/LTE Fe-peak relative abundances $[Sc/Fe]$, $[Cr/Fe]$, and $[Co/Fe]$ versus $[Fe/H]$ (Data from the compilation of Frebel 2010). See text for discussion

$[Fe/H] \sim -3.5$ to roughly solar values at higher metallicities ($[Fe/H] \sim -1.0$). Finally, Sc and Ni remain relatively unchanged with respect to $[Fe/H]$.

Investigations of these very different behaviors have involved two essentially different approaches. The first was to consider whether they could be explained in terms of the explosion energy and position of the mass cut of core-collapse SN models (see Umeda and Nomoto 2005 and references therein), which has been only partially successful and to which the reader is referred. Second, in the context of non-LTE effects, Bergemann and coworkers (e.g., Bergemann et al. 2010) have reported that for each of Cr I, Mn I, and Co I, abundance differences (in the sense $\Delta[X/Fe](\text{non-LTE} - \text{LTE})$) are small at high metallicity and increase to $\sim +0.4$ to $+0.6$ at $[Fe/H] \sim -3.0$. Consideration of ● Figs. 3-11 and ● 3-15 shows that while this acts to remove the downward trends for $[Cr\ I/Fe]$ and $[Mn/Fe]$ (and consistency then

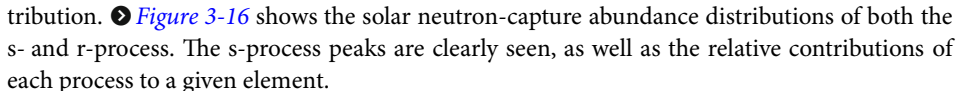
with [Cr II/Fe] results), it exacerbates the upward behavior observed for [Co/Fe], leading to an excess of 1 dex above the solar value at [Fe/H] = -3.5 – providing an even larger challenge for an understanding of this abundance ratio.

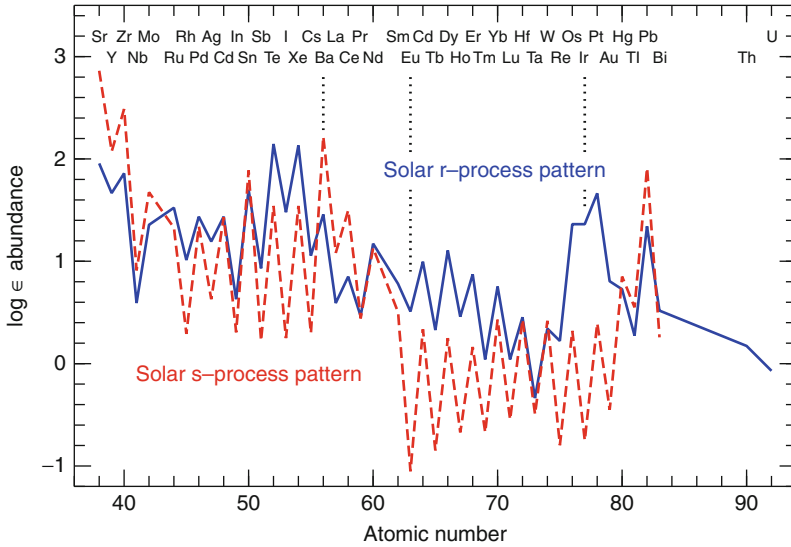
4.2.2 The Evolution of Neutron-Capture Elements

Only elements up to zinc can be synthesized via nuclear fusion. Most heavier elements are built up by the slow neutron-capture process, the *s*-process, and the rapid neutron-capture process, the *r*-process (e.g., Meyer 1994 and references therein). The *s*-process operates under conditions of relatively low neutron densities ($n_n \sim 10^7$ neutrons cm^{-3}). In this regime, the timescale for neutron-capture is slower, in general, than the β -decay rate of unstable isotopes. In contradistinction, when an extremely strong neutron flux is present, the *r*-process occurs on timescales of only a few seconds so that neutron-capture takes place within the β -decay rates of the newly created unstable isotopes. The majority of elements with $Z > 30$ can be produced by either the *s*- or *r*-process, and it is not trivial to disentangle the different production mechanisms. Metal-poor stars offer an opportunity to obtain “clean” nucleosynthetic signatures of each process, as will be described in this section. This opportunity provides unparalleled insight into the details of nucleosynthesis in the early Universe and the onset of chemical evolution of the heaviest elements.

s-process

s-process nuclei are produced in the interiors of low- and intermediate-mass AGB stars and in the He- and C-burning phases of massive stars. On timescales longer than that of a β -decay, neutrons are added to a seed nucleus (i.e., Fe) to buildup heavier, stable nuclei. When neutron-capture creates a radioactive isotope, it will in general decay to its stable daughter isotope before capturing another neutron. In this way, nuclei along the “valley of β -stability” are created. The overall extent to which heavier and heavier isotopes are made is determined by the strength of the neutron flux and the timescale over which it operates. This is known as the time-integrated neutron-flux or neutron-exposure. As a consequence, the *s*-process is more efficient in low metallicity AGB stars due to a relatively larger ratio of neutrons to Fe seeds owing to the primary nature of the neutron source. In massive stars, however, the efficiency of the *s*-process strongly depends on whether the neutron source is of primary or secondary nature and may depend on stellar rotation.

About half of the isotopes of the elements heavier than iron can be created through the *s*-process. Nuclei with atomic numbers that are equivalent to the magic numbers of neutrons, $A = 90$ ($N = 50$), $A = 140$ ($N = 82$), $A = 208$ ($N = 126$), are produced in larger quantities owing to their small neutron-capture cross sections. (Here and in what follows, A refers to the mass number, Z to the atomic number, and N to the neutron number of a nucleus.) This results in three so-called *s*-process peaks that make up a distinct neutron-capture abundance signature. The first peak is located at Sr, Y, and Zr; the second at Ba, La, and Ce; and the third occurs at the end point of the *s*-process, Pb and Bi. For the Sun, the *s*-process component can be calculated and subtracted from the total neutron-capture pattern to obtain the *r*-process contribution.  **Figure 3-16** shows the solar neutron-capture abundance distributions of both the *s*- and *r*-process. The *s*-process peaks are clearly seen, as well as the relative contributions of each process to a given element.



■ Fig. 3-16

Solar *s*- and *r*-process patterns (*dashed* and *full* lines, respectively) (data from Burris et al. 2000). Even after billions of years of chemical evolution, the different contributions of the *s*- and *r*-process still have distinct patterns. Sr and Ba are predominantly produced in the *s*-process, whereas Eu, Os, Ir, and Pt originate mainly in the *r*-process (The vertical dotted lines have been added to facilitate identification of Ba, Eu, and Ir)

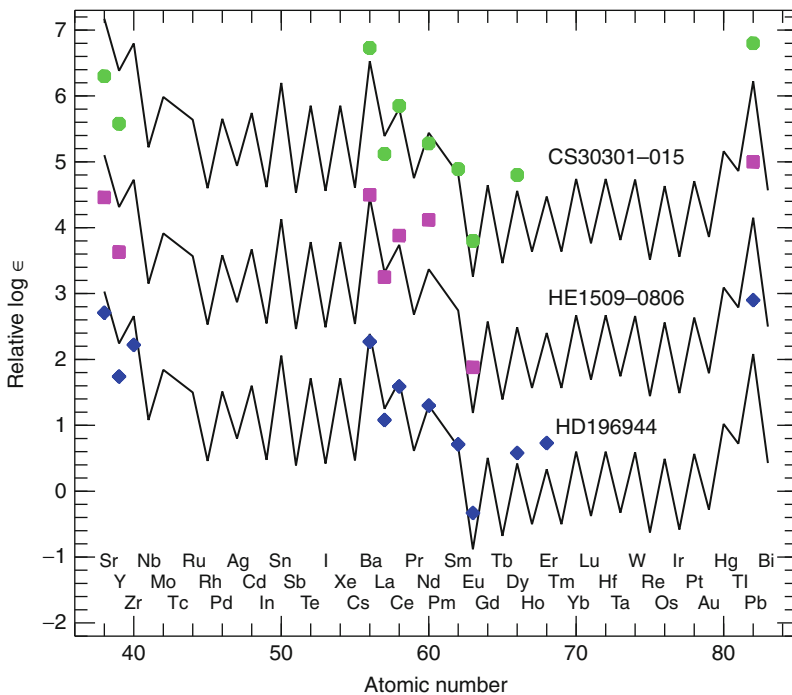
Overall, the *s*-process is rather well understood theoretically, even though there remain uncertainties with regard to the modeling of the amount of ^{13}C that acts as a major neutron source and other reaction rates associated with it (e.g., Arlandini et al. 1999; Sneden et al. 2008). The “main” component of the *s*-process occurs in the helium shells of thermally pulsing lower mass AGB stars and is believed to account for elements with $Z \geq 40$. Examples of *s*-process yields obtained from models of AGB stars with masses of 1–6 M_{\odot} , and different metallicities can be found in Karakas and Lattanzio (2007).

The main neutron sources are α -captures on ^{13}C and ^{22}Ne nuclei. The former creates a low neutron density of $n_n \sim 10^7$ neutrons cm^{-3} , whereas the latter can provide a burst of neutrons with fluxes up to $n_n \sim 10^{13}$ neutrons cm^{-3} during convective thermal pulses. The concentration of ^{13}C and the low reaction rate at the temperatures under which the $^{13}\text{C}(\alpha, n)^{16}\text{O}$ reaction occurs in the He shell maintain the *s*-process for thousands of years. Moreover, the repeated exposure of the He shell to neutron fluxes is important for forming the heaviest elements in AGB stars. On the contrary, the $^{22}\text{Ne}(\alpha, n)^{25}\text{Mg}$ source has a timescale of only ~ 10 years. During the final stages of AGB evolution, *s*-process material is dispersed by stellar winds.

The so-called weak component of the *s*-process occurs in the He- and C-burning cores of more massive stars of roughly solar metallicity and preferentially produces elements around $Z \sim 40$. These stars are just massive enough (perhaps around 8 M_{\odot}) to eventually explode as core-collapse SNe during which the *s*-process material is ejected into the ISM. Regardless of the mass range, the AGB phase includes a series of dredge-up episodes that transport the newly

created material to the surface. Through stellar winds, the ISM is immediately enriched with s-process elements, making AGB stars significant contributors to galactic chemical evolution.

Given that many stars occur in binary systems, a common scenario is mass transfer during which s-process elements are transferred to a lower mass companion. This fortuitously provides an indirect method of studying a clean AGB nucleosynthesis signature. The process occurs not only in the early Universe, but also among higher metallicity stars. The so-called Ba stars are the “receiver” stars within Population I binaries and the “CH stars” those within mild Population II systems. The characteristic s-process signature seen in [Fig. 3-16](#) has been observed in many metal-poor stars as the result of a more massive companion going through the AGB phase and transferring some material onto its companion (e.g., Aoki et al. 2001). In [Fig. 3-17](#), abundances for several s-process-enhanced stars are shown in comparison with the scaled-solar s-process pattern. (One should recall here that an abundance definition that classifies stars as s-process-rich is given in [Table 3-1](#).) As may be seen in the figure, there is good agreement between the scaled solar-pattern and the stellar abundances. (It should be remarked in passing that the relatively poor agreement for Pb results from a significant underproduction of Pb in earlier s-process solar models, such as the one presented in [Fig. 3-17](#) (R. Gallino private communication). Correspondingly, the scaled-solar r-process Pb predictions are too high. There



■ Fig. 3-17

1D/LTE abundances of s-process-enhanced metal-poor stars compared with the scaled-solar s-process pattern. See text for discussion

remain disagreements, however, between the observed Pb abundances and the model predictions, suggesting that either our understanding of these processes is still rather limited or that there are systematic uncertainties in the abundance determinations, or both.)

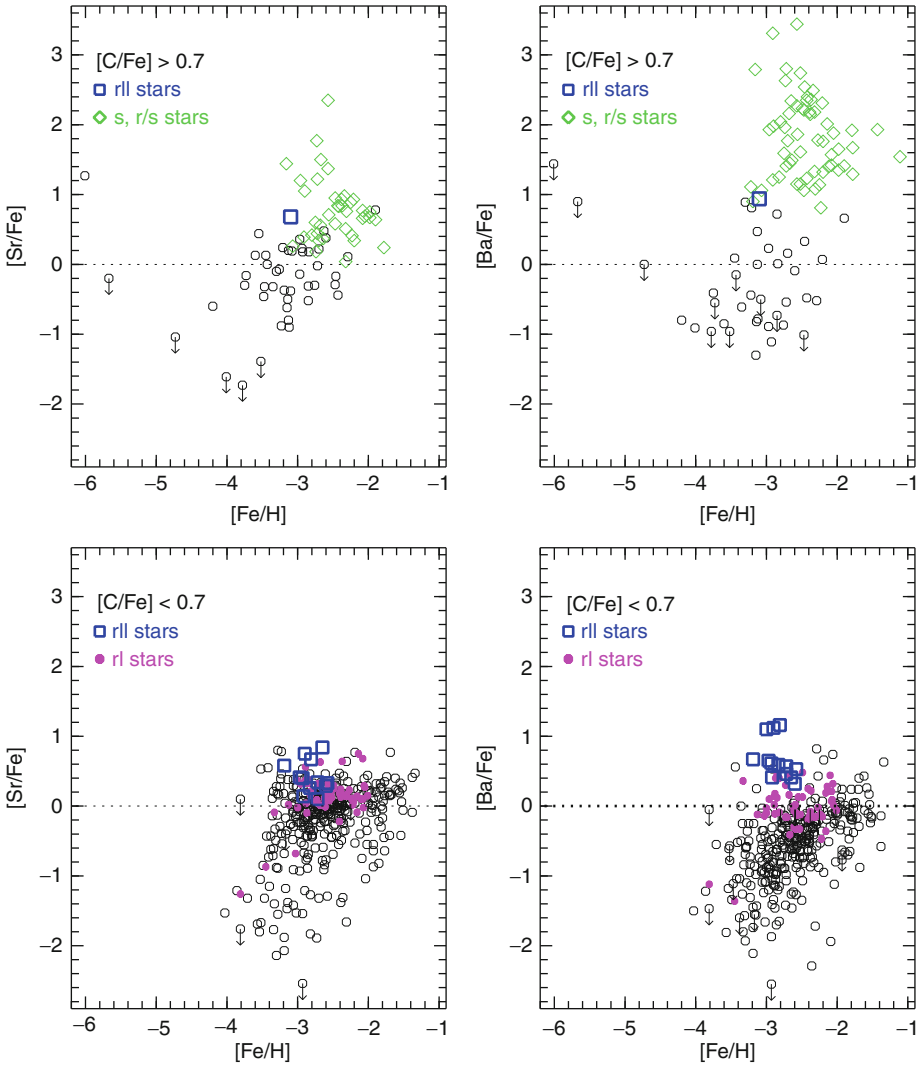
This agreement is remarkable given the fact that the solar neutron-capture material is a product of ~ 8 Gyr of integrated chemical evolution, whereas the halo stars received these elements directly from one of the AGB stars that made s-process elements early in the Universe. Overall, the abundance match indicates a solid theoretical understanding of the s-process. This is also demonstrated by a small number of s-process stars that show extremely large enhancements of Pb, as predicted for third peak elements (see Sneden et al. 2008). Additionally, since carbon is also produced during the AGB phase, the mass transfer usually includes large amounts of carbon. (s-process-enhanced stars are marked as such in [Fig. 3-13](#), with diamond symbols, to illustrate this point.) Most importantly, it should be kept in mind that the carbon excess in these stars is dominated by an extrinsic source and not representative of the intrinsic carbon abundance of the stars' birth cloud. Finally, note that despite the mass transfer, the s-process-enhanced metal-poor stars exhibit lighter element ($Z < 30$) abundance patterns that are the same as those of other normal metal-poor halo stars. One exception is fluorine which, if significantly enhanced, is a signature of low-mass AGB pollution together with the usual s-process-element and carbon-enhancements.

The evolution of representative neutron-capture elements as a function of $[\text{Fe}/\text{H}]$ is shown in [Fig. 3-18](#). Sr and Ba are predominantly produced in the s-process (89% and 85% in the Sun, respectively); see Burris et al. (2000) for details. Since the first lower mass stars in the Universe reached their AGB phase ~ 1 Gyr after the Big Bang, s-process enrichment occurs with some delay with respect to core-collapse SN enrichment. This is indeed as observed: at a metallicity of $[\text{Fe}/\text{H}] \sim -2.6$, the s-process is in full operation, including significant neutron-capture-element "pollution" of the galaxy by AGB stars (Simmerer et al. 2004), as can be seen in the top panels of [Fig. 3-18](#). Metal-poor stars with an obvious s-process signature from a mass transfer event are also carbon-rich. At these and higher metallicities, all stars thus formed from gas that had already been enriched in s-process elements, irrespective of whether or not they received extra s-process material from a companion. As can be seen in the bottom panels of the figure, there is a main branch in the $[\text{Ba}/\text{Fe}]$ versus $[\text{Fe}/\text{H}]$ plane. Above $[\text{Fe}/\text{H}] \sim -2.6$, it is dominated by stars formed from AGB-enriched gas.

There are some exceptions with respect to clean s-process signatures in metal-poor stars. A handful of objects display a mixed abundance signature originating from both the s- and the r-process (see, e.g., Jonsell et al. 2006 for an extensive discussion on the origin of s-/r-mixtures observed in metal-poor stars). This includes some stars with $[\text{Fe}/\text{H}] < -2.6$, and their unusual chemical patterns are perhaps due to earlier more massive stars expelling some s-process elements when they exploded as core-collapse SNe. Several different scenarios have been invoked to explain the combination of the two neutron-capture processes originating at two very different astrophysical sites. No completely satisfactory explanation, however, has yet been found.

r-process

Heavy elements are also produced in the rapid (r-) process, which takes place over just a few seconds. Seed nuclei (e.g., C or Fe) are bombarded with neutrons ($\sim 10^{22}$ neutrons $\text{cm}^{-2} \text{s}^{-1}$) to quickly form large radioactive nuclei far from stability. After the strong neutron flux ceases, the nuclei decay to form stable, neutron-rich isotopes. The r-process does not, however, produce infinitely large nuclei because of a significant decrease in the cross sections of neutron-capture



■ Fig. 3-18

1D/LTE neutron-capture-element abundances ratios $[Sr/Fe]$ and $[Ba/Fe]$ as a function of $[Fe/H]$ for carbon-enhanced objects ($[C/Fe] \geq 0.7$; *top panels*) and other halo stars (*bottom panels*). The range in $[Sr/Ba]$ and $[Ba/Fe]$ is much larger than uncertainties or systematic differences between individual studies, indicating a cosmic origin. Below $[Fe/H] \sim -3.0$, the evolution is dominated by r-process enrichment. For $[Fe/H] \gtrsim -2.6$, the s-process significantly contributes neutron-capture material (see Simmerer et al. 2004). Arrows indicate upper limits, while the solar ratio is indicated by dotted lines

nuclei with closed neutron shells. Other unfavorable reaction rates and problems with nuclear stability in the heavy-isotope region also play a role. These factors eventually terminate the r-process at nuclei around $A = 270$, far in the transuranium regime. Those nuclei all decay to eventually become Pb. Approximately half of the neutron-capture isotopes heavier than iron

are produced in this way, including the heaviest, long-lived radioactive elements thorium and uranium.

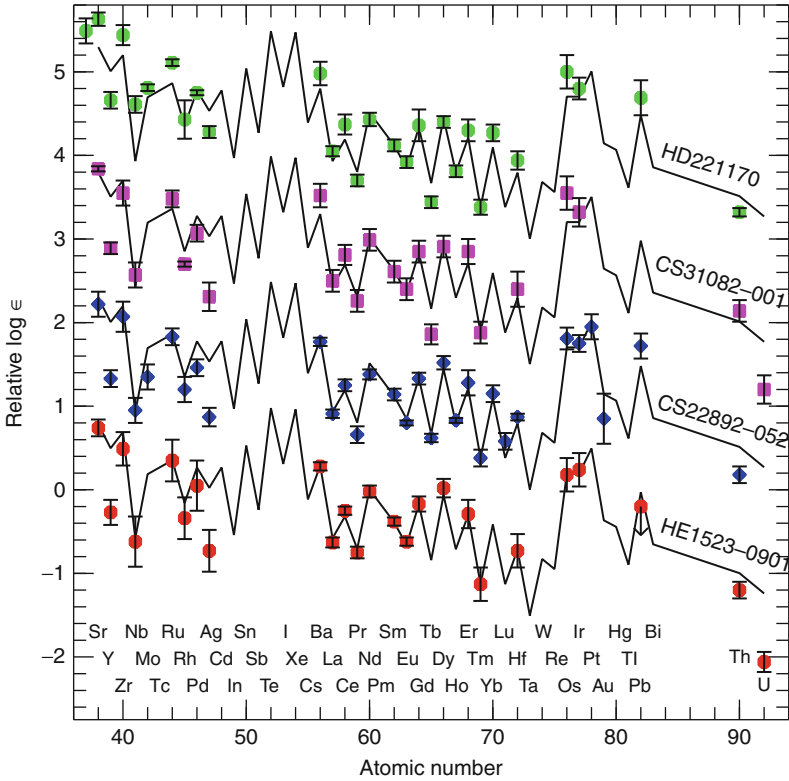
The r-process also manifests itself in a characteristic abundance pattern, showing three large peaks at elements with $A \sim 80$ ($Z \sim 33$; Se-Br-Kr), $A \sim 130$ ($Z \sim 52$; Te-I-Xe), and $A \sim 195$ ($Z \sim 77$; Os-Ir-Pt), similar to the s-process peaks. The latter two of these may be seen in [Fig. 3-16](#). The peaks form because nuclei with closed neutron shells only reluctantly capture any neutrons (i.e., they have extremely small cross sections). With their long β -decay lifetimes, they act as bottlenecks to additional neutron captures creating even heavier nuclei. Hence, nuclei with atomic masses at or just below the closed-shell nuclei pile up during the process.

Unlike the situation for the s-process, the astrophysical site(s) that provide the extreme neutron fluxes required for the r-process have not yet been identified. Neutron-star mergers have been considered, but their long evolutionary timescale prior to merging argues against them being the primary r-process site in the early Galaxy. Neutrino-driven winds emerging from the formation of a neutron star during a core-collapse SN explosion are more promising locations (Qian and Wasserburg 2003). Since massive SNe dominate chemical enrichment in the early Universe (e.g., as documented through the α -element enhancement found in halo stars), the neutrino-driven wind model agrees naturally with such an early SN enrichment mode.

In order to learn about the details of the r-process and its site, it is of great importance to obtain actual data of a clean r-process signature. The best candidates for this are the r-II stars (see [Table 3-1](#) for definitions), the most strongly r-process-enhanced objects, which comprise about 5% of stars with $[\text{Fe}/\text{H}] \lesssim -2.5$ (see Barklem et al. 2005). All but one of the r-II stars have metallicities close to $[\text{Fe}/\text{H}] = -3.0$, with the outlier having an even lower $[\text{Fe}/\text{H}]$. The metallicities are thus distinctly lower than the value of $[\text{Fe}/\text{H}] = -2.6$ discussed above as corresponding to the onset for AGB s-process enrichment. (It should be recalled that mildly enriched r-I stars are found up to metallicities of $[\text{Fe}/\text{H}] \sim -2.0$, while at higher values, the signature becomes less clean since the more metal-rich star would have formed from material already significantly enriched by many more than just one r-process event.) This suggests that the r-process enhancement comes from stars more massive than those that experience the s-process during AGB evolution.

The “main” r-process operates in the full range of neutron-capture elements, up to $Z = 92$. Model calculations have shown that it probably only occurs in a specific, yet unidentified type of core-collapse SN or perhaps only in a particular mass range ($\sim 8\text{--}10 M_{\odot}$; Qian and Wasserburg 2003). Examination of the ratios of the heavy ($Z > 56$) neutron-capture abundances in r-process-enhanced stars (e.g., Barklem et al. 2005; Frebel et al. 2007a; Hill et al. 2002; Sneden et al. 1996) shows that the abundance distribution of each closely matches that of the scaled-solar r-process pattern (e.g., Burris et al. 2000). [Figure 3-19](#) shows data for four well-studied r-II stars. Given that the Sun was born ~ 8 Gyr later than these otherwise ordinary metal-poor stars, this is a remarkable finding. Assuming that the r-process takes place only in core-collapse SNe, the match of the stellar and solar patterns suggests that the r-process is universal: that is, no matter when and where it happens, it always produces its elements with the same proportions. Otherwise, the integrated pattern observed in the Sun would not resemble the individual pattern found in a ~ 13 Gyr old star.

While there is excellent agreement with the scaled-solar r-process pattern for elements heavier than Ba, deviations have been found among the lighter neutron-capture species. This indicates that the origin of the lighter elements is more complex, with perhaps both the “main” and “weak” r-processes contributing in different mass ranges (see, e.g., Travaglio et al. 2004). The “weak” r-process is thought to produce mainly the lighter neutron-capture elements



■ Fig. 3-19

1D/LTE abundances in *r*-process-enhanced metal-poor stars compared with the scaled solar *r*-process pattern. Note the remarkable agreement for elements heavier than Ba ($Z \geq 56$)

($Z < 56$) and little or no heavier material, such as Ba. Possibly, this occurs mainly in massive ($\geq 20 M_{\odot}$) core-collapse SNe (see, e.g., Wanajo and Ishimaru 2006). A candidate for an observed “weak” *r*-process signature is provided by the *r*-process-poor, metal-poor star HD 122563 (Honda et al. 2006), which displays a depleted, exclusively light neutron-capture-element pattern. The [Sr/Ba] ratio in this and other stars can be employed to learn about the relative contributions of the two *r*-processes, and potentially the origin of the overall abundance pattern. In this scenario, the “main” *r*-process would produce lower [Sr/Ba] ratios than the “weak” one.


Overall, neutron-capture elements are produced in limited amounts. Their abundances in solar system material is about six orders of magnitude less than those of the Fe-peak elements. Nevertheless, they provide invaluable insight into various early nucleosynthesis processes. The enormous scatter of neutron-capture abundances (e.g., [Ba/Fe]), as a function of [Fe/H], suggests that the production of neutron-capture elements is completely decoupled from that of Fe and other elements. As described earlier, and displayed in \blacktriangleright Figs. 3-14 and \blacktriangleright 3-15, the α - and Fe-peak element abundances show very little scatter, probably because the ISM was already relatively well mixed by [Fe/H] ~ -4.0 . As can be seen in \blacktriangleright Fig. 3-18, especially at [Fe/H] ~ -3.0 , there is a range of ~ 6 dex in neutron-capture abundances. This must reflect strongly varying degrees of neutron-capture yields at the earliest times and probably also different processes contributing different groups of these elements in various amounts (e.g., more Sr than Ba at the

very lowest metallicities). Only at somewhat higher metallicities, when the s-process begins to dominate the neutron-capture inventory, does the bulk of the stellar abundance ratios become more solar-like.

4.3 The Milky Way Globular Clusters and Dwarf Galaxies

4.3.1 Globular Clusters

The internal relative abundance patterns of the Galactic globular clusters are distinctly different in many respects from those of the Galactic halo and require an explanation involving poorly understood intracluster self-enrichment processes. While this is a very important field of endeavor, insofar as it lies beyond the scope of the present chapter, a comprehensive description of this topic cannot be provided here. Instead, some of the main differences between the globular cluster and halo field stars are briefly highlighted:

- Most clusters are chemically homogeneous with respect to iron, at the ~ 0.03 dex level. The clear exceptions among the halo clusters are ω Centauri, M22, and M54, where ranges of $\Delta[\text{Fe}/\text{H}] \sim 0.3$ to 1.5 dex have been observed. In  Fig. 3-9, the reader can see the MDF of ω Cen, the cluster with the largest spread; a large number (~ 5) of subpopulations, with distinctly different mean $[\text{Fe}/\text{H}]$, have been identified in this system.
- All globular clusters so far studied have been found to be chemically inhomogeneous in a number of light elements that are produced or modified in nucleosynthetic (p, γ) reactions. Beginning in ~ 1970 , observations and analyses of cluster members have over time added the following elements to the list – C, N, O, Mg, Na, and Al. Strong correlations and anticorrelations exist among the abundances of these elements, and within a given cluster, subpopulations have been identified based on abundance patterns involving them. Lithium and heavy-neutron-capture-element variations have also been reported that correlate with those of the above elements in some clusters. Intermediate-mass AGB stars are most commonly identified as the nucleosynthesis sites responsible for these variations, together in some cases with internal mixing high on the present-day RGB.
- Finally, some (but not all) of the most massive clusters (ω Cen, NGC 2808) show multiple main sequences in the color-magnitude diagram, for which the only empirically consistent explanation yet proposed is that there are subpopulations within these systems that have distinct helium abundances in the astoundingly large range of $Y \sim 0.23\text{--}0.35$ (Y is the helium fraction by mass). No completely satisfactory explanation has yet been given, although several authors identify massive AGB stars in an early generation of cluster stars as the prime candidate.

Various models have been proposed that are unique to the globular cluster environment and which involve a number of stellar generations that chemically enrich the material from which subsequent generations form. An example of such a model, which also provides references to the observational material described above, is provided by Conroy and Spergel (2011).

4.3.2 Dwarf Galaxies

One of the important unsolved problems in cosmology is understanding the formation of galaxies. Studying the compositions of stars in dwarf galaxies provides information on the chemical

evolution of these systems. The Milky Way's dwarf galaxy satellites, with a large range in masses and luminosities very different from those of the Milky Way itself, permit a comparison of their chemical evolution histories, which in turn provides clues to the origin and overall evolution of different types of galaxies. Specifically, the connection between the surviving dwarf systems and those believed to have been captured and dissolved to form the Milky Way halo is best addressed by examining in detail the stellar chemical abundances of present-day dwarf galaxies (see also [Sect. 6.2](#) on this topic). The most-metal-poor (and hence oldest) stars in a given system permit unique insight into the earliest phases of star formation. Stars born at later times (and thus with higher metallicities) contain the integrated effects of internal chemical evolution in their atmospheric compositions. (See Kirby et al. 2011 for an overview of the history and current state of simple chemical evolution models.)

Dwarf spheroidal galaxies are relatively simple systems that allow us to study, both observationally and theoretically, the basic processes that led to their origin and evolution. They are generally old, metal-poor, have no gas, and thus no longer support star formation. On the other hand, a large fraction of their mass comprises dark matter, with the least luminous of them having mass-to-light ratios of order 10^3 (in solar units). Some 25 such systems are currently known orbiting the Galaxy today (see Tolstoy et al. 2009 for a review). The ~ 10 recently discovered “ultra-faint” dwarf galaxies ($L_V \leq 10^5 L_\odot$; Martin et al. 2008) are some orders of magnitude fainter than the more luminous, “classical,” Milky Way dwarf spheroidal galaxies. As has been outlined in [Sect. 3.2.1](#), all of these dwarf systems follow a (metallicity, luminosity) relationship, with the classical dwarfs being on average more metal-rich and containing more stars than their less luminous ultra-faint siblings.

With $[\text{Fe}/\text{H}] \gtrsim -2.0$, stars in the classical dwarf galaxies were found to have abundance ratios different from halo stars at the same metallicity (e.g., Geisler et al. 2005; Shetrone et al. 2003). Most strikingly, the α -element abundances are not enhanced to the SN II enrichment level of $[\alpha/\text{Fe}] \sim 0.4$. This indicates different enrichment mechanisms and longer timescales in the dwarf galaxies; due to a slower evolution, the Fe contribution from SN Ia occurred “earlier,” at a time when the entire system had not yet reached a metallicity of $[\text{Fe}/\text{H}] \sim -1.0$, the turndown point of $[\alpha/\text{Fe}]$ versus $[\text{Fe}/\text{H}]$ in the Milky Way (see [Fig. 3-14](#)).

Only very recently, a handful of stars with metallicities of $[\text{Fe}/\text{H}] < -3.0$ were discovered in the classical dwarf galaxies, with some of them having $[\text{Fe}/\text{H}] \sim -4.0$ (e.g., Frebel et al. 2010a). While these dwarfs have been studied for many decades, problems with earlier search techniques had prevented the discovery of extremely metal-poor stars (Starkenburger et al. 2010). The existence of such objects shows that a metallicity range of ~ 3 dex is present, at least in the Sculptor and Fornax dSphs. At $[\text{Fe}/\text{H}] < -3.0$, the chemical abundances, obtained from high-resolution spectra, are remarkably similar to those of Galactic halo stars at similar metallicities. This is in contrast to the deviations at higher $[\text{Fe}/\text{H}]$ and provides evidence for a change in the dominant enrichment mechanisms. For these types of dwarf galaxies, the transition from halo-typical abundance ratios (as a result of SN II enrichment) to more solar-like values (SN Ia-dominated Fe production) appears to take place around $[\text{Fe}/\text{H}] = -3.0$ (Aoki et al. 2009; Cohen and Huang 2009). As a consequence, chemical evolution may be a universal process, at least at the earliest times, the very regime that is probed by the most-metal-poor stars.

The first extremely metal-poor stars not belonging to the Galactic halo field population were found in some of the ultra-faint dwarf galaxies, even before such stars were discovered in the classical dwarfs (Kirby et al. 2008). Due to their distance and low stellar density, these systems contain few stars brighter than $V = 19$, making the collection of spectroscopic data a challenge. Nevertheless, high-resolution spectra of a handful of individual metal-poor stars in Ursa Major II, Coma Berenices, Bootes I, Segue 1, and Leo IV (Frebel et al. 2010b;

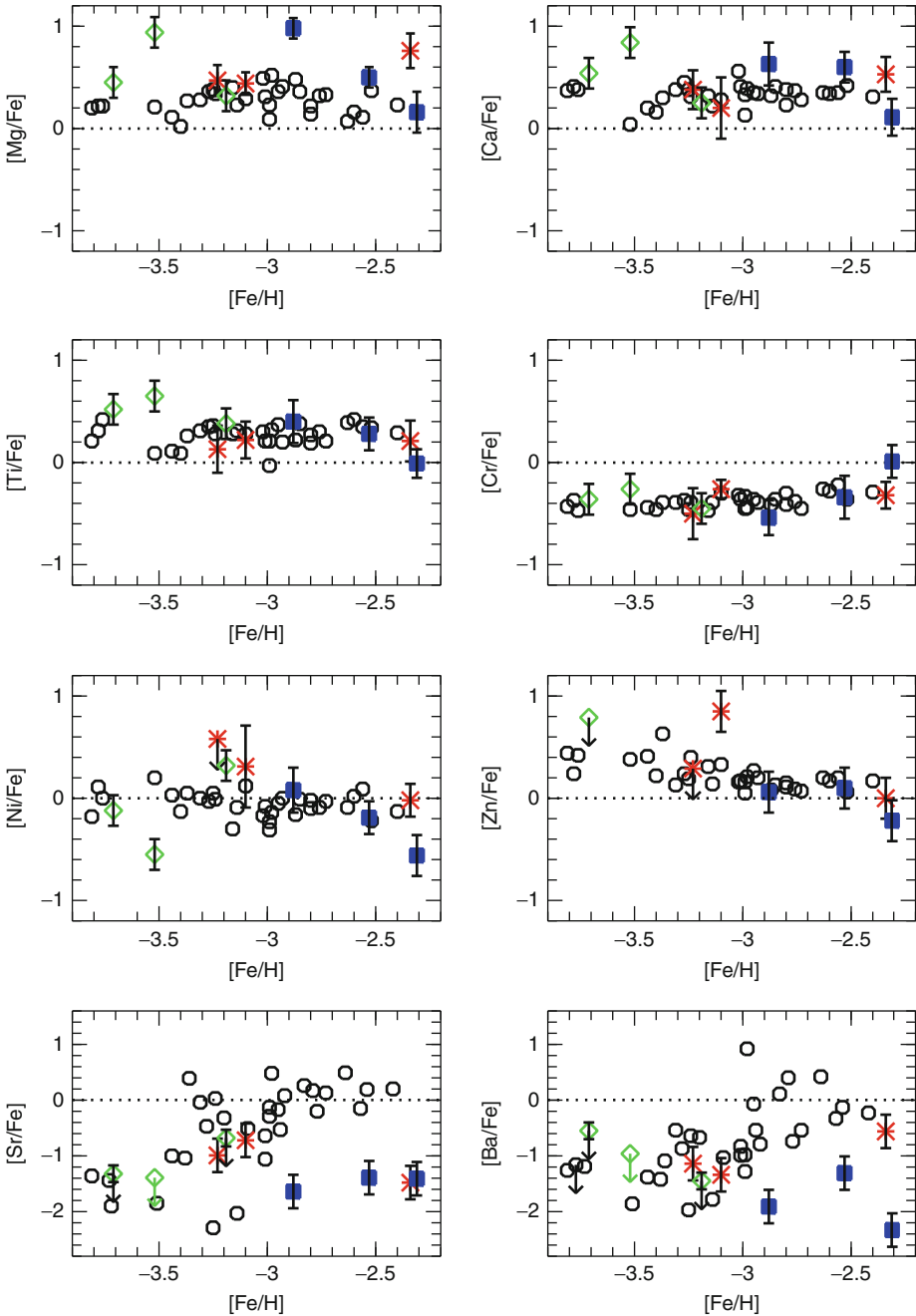
Norris et al. 2010a, c; Simon et al. 2010) have been obtained. A large fraction of them are extremely metal-poor (i.e., $[\text{Fe}/\text{H}] < -3.0$). With one exception, all of their light element ($Z < 30$) abundances show the halo-typical core-collapse SNe signature, resembling those of similarly metal-poor Galactic halo stars. This may be clearly seen in the upper six panels of [Fig. 3-20](#), where the relative abundances of the metal-poor halo red giant sample of Cayrel et al. (2004) and François et al. (2007) (presented above in [Sect. 3.2.2](#)) are compared with those available for red giants in the ultra-faint dwarf galaxy stars. The exception is the CEMP-no star (see [Table 3-1](#)) Segue 1-7, a radial-velocity member of Segue 1, which has $[\text{Fe}/\text{H}] = -3.5$, $[\text{C}/\text{Fe}] = +2.3$, and $[\text{Ba}/\text{Fe}] < -1.0$ (Norris et al. 2010a). The 200-fold overabundance of carbon relative to iron in this extremely metal-poor star is quite remarkable. This shows that the CEMP-no phenomenon is not restricted to the Milky Way halo and may provide important clues to the origin of these stars.

Some abundances, however, indicate that the chemical evolution in these small systems may have been moderately inhomogeneous (see [Sect. 3.2.1](#), [Fig. 3-9](#)), with some stars perhaps reflecting enrichment by massive Population III stars. The chemical similarity to halo stars is also found at higher metallicity, up to $[\text{Fe}/\text{H}] \sim -2.0$, in contrast to what has been found for the classical dwarfs. This remarkable similarity between the abundance profiles of the halo and the dwarf galaxies supports the view that chemical evolution is independent of galaxy host mass in this metallicity regime. Moreover, this (together with the existence noted above of a CEMP-no star in the ultra-faint Segue 1) renews support for a scenario in which the metal-poor end of the Milky Way halo population was built up from destroyed dwarf galaxies (see [Sect. 6.2](#)).

Finally, neutron-capture abundances should be mentioned. These are extremely low in the ultra-faint systems, and as may be seen in the two bottom panels of [Fig. 3-20](#) in the range $-3.0 < [\text{Fe}/\text{H}] < -2.0$, the observed Sr and Ba values lie well below those found in typical Milky Way halo stars. A more general statement is that the mean values of $[\text{Sr}/\text{Fe}]$ and $[\text{Ba}/\text{Fe}]$ are significantly smaller in the ultra-faint dwarfs than in the halo. Comparably low values for Sr and Ba are also found in the more luminous dwarfs Hercules (Koch et al. 2008) and Draco (Fulbright et al. 2004) despite their sometimes relatively high Fe values of $[\text{Fe}/\text{H}] \sim -2.0$.

5 Cosmo-Chronometry

Because of their low metallicity, metal-poor stars are usually regarded as having formed at the earliest times, when the first elements heavier than helium were being synthesized. The most metal-poor stars are thus regarded as being almost as old as the Universe. Age determinations for field stars are, however, difficult, since they do not belong to a distinct single-age population such as a globular cluster. Cluster ages are based on fitting isochrones to their color-magnitude diagrams. The age dating of globular clusters will not be discussed here, and the reader is referred to Vandenberg et al. (1996) for details and to Marín-Franch et al. (2009) for more recent results which are addressed further in [Sect. 6.2](#). Suffice it to say that although the clusters are not as metal-poor as the most metal-poor field stars, the ages of the older of them range from 10 to 14 Gyr, placing them among the oldest objects in the Universe. The main point of focus in this chapter is dating techniques for individual r-process-enhanced Galactic halo field stars.



■ Fig. 3-20

Comparison of stars in the Galactic halo (*circles*: Cayrel et al. 2004; François et al. 2007) and dwarf galaxies (*asterisks*: Ursa Major II, *filled squares*: Coma Berenices, *diamonds*: Segue 1, Bootes I and Leo IV) in the 1D/LTE relative abundances ($[X/Fe]$) versus $[Fe/H]$ diagram. While the light element abundances agree very well, dwarf galaxy stars have relatively low neutron-capture abundances, albeit still within the range of the halo stars

5.1 Nucleo-chronometry of Metal-Poor Field Stars

A fundamental way to determine the age of a *single* star is through radioactive decay dating. Elements suitable for this procedure are not, however, present in sufficient quantities in ordinary stars. There is also the problem of finding stars that have experienced enrichment from a single source so that the decay tracks the time from just one production event until the time of measurement. Fortunately, in strongly r-process-enhanced metal-poor stars, radioactive age dating is possible through abundance measurements of Th (^{232}Th , half-life 14 Gyr) and/or U (^{238}U , half-life 4.5 Gyr). These half-lives are sufficiently long for measurements of cosmic timescales, and stellar ages can be determined based on radioactive decay laws that lead to simple equations for different chronometer ratios involving Th, U, and stable r-process elements. Observed abundances in r-process-enhanced stars provide determinations of their remaining radioactive material, for example, $\log \epsilon(\text{Th}/r)_{\text{now}}$, with r being a stable element such as Eu, Os, and Ir, for which the following relationships (derived from radioactive decay laws in combination with known nuclear physics) obtain:

1. $\Delta t = 46.78[\log(\text{Th}/r)_{\text{initial}} - \log \epsilon(\text{Th}/r)_{\text{now}}] \text{ Gyr}$
2. $\Delta t = 14.84[\log(\text{U}/r)_{\text{initial}} - \log \epsilon(\text{U}/r)_{\text{now}}] \text{ Gyr}$
3. $\Delta t = 21.76[\log(\text{U}/\text{Th})_{\text{initial}} - \log \epsilon(\text{U}/\text{Th})_{\text{now}}] \text{ Gyr}$

Only theoretical r-process calculations can provide the initial production ratios ($\log(\text{Th}/r)_{\text{initial}}$ and $\log(\text{U}/r)_{\text{initial}}$) that describe how much r-process material, including Th and U, was made in the production event, i.e., the SN explosion. This implies that, technically, the SN is dated rather than the star. The time span, however, of the formation of the star after the SN is regarded as negligibly short compared to the star's age. Currently, the astrophysical site of the r-process remains unclear, and the associated initial conditions are not known, making yield predictions difficult. Nevertheless, some calculations involving various approximations are available (e.g., Schatz et al. 2002). It should be kept in mind that the universality of the r-process, noted in [► Sect. 4.2.2](#), (at least for $Z \geq 56$) is a major ingredient in predicting the relative elemental ratios, such as Th/ r .

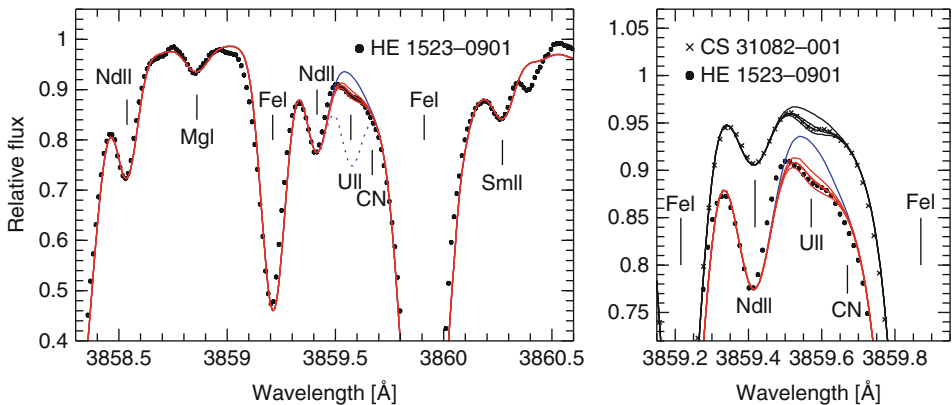
The first r-II star, CS 22892-052, was discovered more than a decade ago in the HK survey (McWilliam et al. 1995). Its Th/Eu ratio yielded an age of 14 Gyr (Snedden et al. 2003). A second object, CS 31082-001, in which both Th and U were measurable, was also shown to be 14 Gyr old, based on its U/Th abundance ratio (Hill et al. 2002). A large campaign was then initiated to observe metal-poor candidate stars from the Hamburg/ESO survey to more systematically discover such objects. Identifying stars with a strong Eu II line at 4,129 Å, the main r-process indicator in stellar spectra, led to the discovery of several strongly r-process-enhanced (r-II) stars (see Barklem et al. 2005, and references therein) and dozens of mildly enriched (r-I) objects. With the exception of CS 31082-001, many of these r-process-enhanced stars could be dated with only the Th/Eu chronometer.

In CS 31082-001, chronometer ratios involving any stable elements (e.g., Th/Eu) yielded *negative* ages. This is due to unusually high Th and U abundances compared with values expected for these elements from their overall r-process pattern (scaled to the Sun). Since only the elements heavier than the 3rd r-process peak are (equally) affected (Roederer et al. 2009), the U/Th ratio still gives a reasonable age for this star. The behavior was termed an “actinide boost” (Schatz et al. 2002) and indicates an origin different from “normal” r-process-enhanced stars and/or multiple r-process sites (Hill et al. 2002). Since then, three more r-process-enhanced stars with such high Th/Eu ratios ($\sim 20\%$ of r-process stars) have been found (Honda et al. 2004;

Lai et al. 2008). The underlying physical process(es) leading to the large fraction of the actinide-boost stars will need to be thoroughly investigated over the next few years. It is crucial to assess whether the apparent universality of the r-process of elements with $Z \geq 56$ seen in “regular” r-process-enhanced stars remains truly universal or if it is simply an artefact of our limited understanding of the r-process and/or insufficient numbers of such stars.

For only one r-II star has it so far been possible to determine ages from more than just one chronometer ratio. The bright giant HE 1523–0901 ($V = 11.1$) has one of the strongest enhancements in r-process elements so far observed, $[r/Fe] = 1.8$ (Frebel et al. 2007a), and among the measured neutron-capture elements are Os, Ir, Th, and U. Its average stellar age of 13.2 Gyr is based on seven chronometers Th/r, U/r, and U/Th involving combinations of Eu, Os, Ir, Th, and U. Only in cool r-II red giants can the many weak and often partially blended neutron-capture features be measured. The two most challenging examples are the extremely weak U II line at 3,859 Å and the even weaker Pb I line at 4,057 Å; these two lines are the strongest optical transitions of the two elements. (It should be mentioned that both lines are blended with a strong CH feature. Hence, U and Pb can be detected best in stars with subsolar carbon abundances, which minimizes the blending effect. In CS 22892-052, a carbon-rich r-process-enhanced star, neither element will ever be measurable.)

► *Figure 3-21* shows the spectral region around the U line in HE 1523–0901. To be useful for age determination, r-II stars should be as bright as possible (preferably $V < 13$) so that very high-resolution spectra with very high S/N can be collected in reasonable observing times. A successful U measurement requires a high-resolution spectrum ($R > 60,000$) with S/N of at least 350 per pixel at 3,900 Å. A Pb measurement may be attempted in a spectrum with $S/N \sim 500$ at 4,000 Å. Only *three* stars have had U measurements. They are HE 1523–0901,



■ Fig. 3-21

Spectrum synthesis of the U line region at 3,860 Å in HE 1523–0901 (*left panel, whole region; right panel, detailed view of just the line*) and also CS 31082-001 (*right panel only*). Dots indicate the observed spectrum and continuous lines present synthetic spectra computed with a range of U abundances for comparison with the observed one. The latter are best illustrated in the *right panel* where the lowest three lines correspond to $\log \epsilon(U) = -1.96, -2.06,$ and -2.16 , and the uppermost line includes no U. The *dotted line* in the *left panel* represents a synthetic spectrum with an estimated U abundance if U was not radioactive and had not decayed over the past ~ 13 Gyr (From Frebel et al. 2007a)

CS 31082-001, and a somewhat uncertain detection in BD +17° 3248, of which the age of HE 1523-0901 is currently the most reliable.

Compared with Th/Eu, the U/Th ratio is more robust against uncertainties in the theoretically derived production ratio because Th and U have similar atomic masses (for which uncertainties largely cancel out; e.g., Wanajo et al. 2002). Hence, stars displaying Th and U are the best for age determination. For the same reason, stable elements of the 3rd r-process peak ($76 \leq Z \leq 78$) are best used in combination with Th and U. Nevertheless, realistic age uncertainties range from ~ 2 to ~ 5 Gyr depending on the chronometer ratio (see Schatz et al. 2002, and Frebel et al. 2007a for discussions). In any case, age measurements of old stars naturally provide an important independent lower limit to the age of the Universe, currently inferred to be $13.73^{+0.16}_{-0.15}$ Gyr with WMAP (Spergel et al. 2007). In the absence of an age-metallicity relationship for field halo stars, the nucleo-chronometric ages thus demonstrate that these metal-deficient stars, with $[\text{Fe}/\text{H}] \sim -3$, are indeed very ancient, leading to the corollary that stars of similar $[\text{Fe}/\text{H}]$, but with no overabundance in r-process elements, have a similar age.

The r-process-enhanced stars fortuitously bring together astrophysics and nuclear physics by acting as a “cosmic laboratory” for both fields of study. They provide crucial experimental data on heavy-element production that are not accessible to nuclear physics experiments. Since different r-process models often yield different final r-process abundance distributions, particularly in the heavy mass range, self-consistency constraints are very valuable. The stellar abundance triumvirate of Th, U, and Pb provides such constraints. These three elements are intimately coupled not only with each other but also to the conditions (and potentially also the environment) of the r-process. Pb is the end product of all decay chains in the mass region between Pb and the onset of dominant spontaneous fission above Th and U. It is also built up from the decay of Th and U isotopes. All three measurements thus provide important constraints on the poorly understood decay channels. They offer an opportunity to improve r-process models which, in turn, facilitates the determination of improved initial production ratios necessary for the stellar age dating.

6 Cosmogony

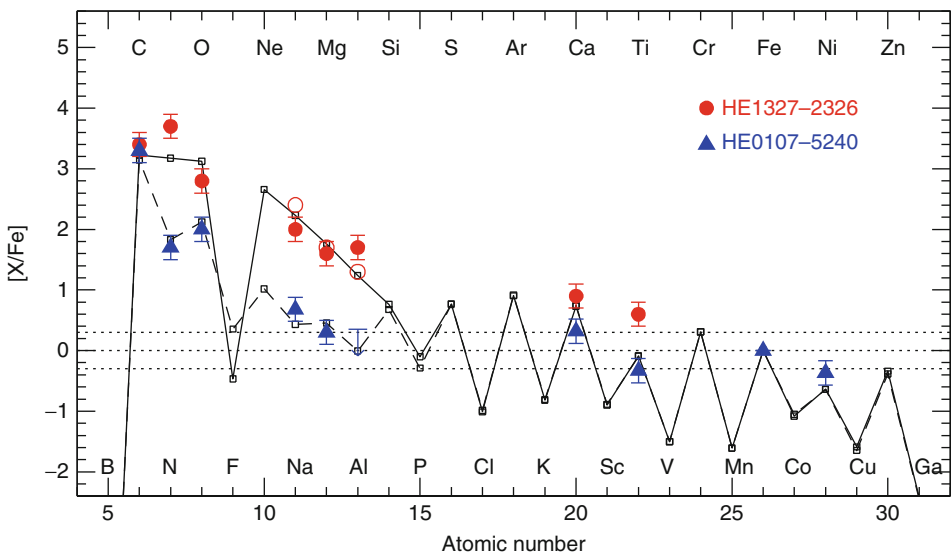
6.1 The Early Universe

Simulations of the hierarchical assembly of galaxies within the cold dark matter (CDM) paradigm pioneered by White and Rees (1978) and today referred to as Λ CDM (e.g., Diemand et al. 2007; Springel et al. 2008) demonstrate that structure formation in the Universe proceeded hierarchically, with small dark-matter halos merging to form larger ones which eventually led to the build-up of larger galaxies like the Milky Way. This is further described in [Sect. 6.2](#). The very first stars (Population III) formed in small, so-called minihalos of $\sim 10^6 M_{\odot}$ that collapsed at $z \simeq 20\text{--}30$ (Tegmark et al. 1997) a few hundred million years after the Big Bang. Hydrodynamical cosmological simulations have shown that due to the lack of cooling agents in primordial gas, significant fragmentation was largely suppressed so that these first objects were very massive, of order $\sim 100 M_{\odot}$ (a “top-heavy” initial mass function; e.g. Bromm and Larson 2004 and references therein), and likely fast rotating. This is in contrast to low-mass stars ($< 1 M_{\odot}$) dominating today’s mass function (often referred to as the Salpeter mass function).

In this scenario, the massive first generation stars (designated Population III.1) soon exploded as core-collapse SNe leaving remnant black holes (for progenitor masses of $25 M_{\odot} < M < 140 M_{\odot}$ and $M > 260 M_{\odot}$) or even more energetic pair-instability SNe (PISNe; $140 M_{\odot} < M < 260 M_{\odot}$; Heger and Woosley 2002) with complete disruption. A specific, predicted “chemical fingerprint” of the putative PISN explosion has not (yet) been identified in any metal-poor star. Given that luminous supernovae (having peak $M_V < -21$) in external galaxies have been associated with massive progenitor stars ($M > 100 M_{\odot}$) in low-metallicity regions (Neill et al. 2011), it cannot be excluded, however, that such a signature will be found. In their final stages, all these massive objects provided vast amounts of ionizing radiation (and some of the first metals) that changed the conditions of the surrounding material for subsequent star formation, even in neighboring minihalos. Partially ionized primordial gas supported the formation of first H_2 and then HD, which in turn facilitated more effective cooling than would be possible in neutral gas. Any metals or dust grains left behind from PISNe would have similar cooling effects. Hence, there likely was a second generation of metal-free stars (Population III.2) that, for the first time, included stars of somewhat smaller masses ($M \sim 10 M_{\odot}$.) This generation, however, was still top heavy, in contrast to typical present-day stars ($M \sim 1 M_{\odot}$). For a recent review of this topic, see Bromm et al. (2009). Soon thereafter, the first low-mass metal-poor stars were born. In their atmospheres, they locked in the chemical fingerprint of the very first supernova explosions. Investigating the chemical abundances of the most metal-poor stars is thus the only way to gain detailed information of the nature and properties of the first stars without going to the very high redshift Universe. Even with the James Webb Space Telescope (see <http://www.jwst.nasa.gov>), the sensitivity will not be sufficient to directly observe the first stars. The first galaxies may, however, just be reachable.

As described in [Sect. 2.4](#), four halo stars with the exceptionally low values of $[Fe/H] < -4.3$ are currently known. An immediate question arises: do their abundance patterns reflect the chemical yields of the first stars? Before attempting to answer this question, their detailed chemical abundances have to be considered. The most striking features in both stars with $[Fe/H] < -5.0$ are the extremely large overabundances of the CNO elements ($[C, N, O/Fe] \sim +2$ to $+4$). HE 0557–4840 (with $[Fe/H] = -4.8$) partially shares this signature by also having a fairly large value of $[C/Fe]$. SDSS J102915+172927, however, does not. This object has an abundance signature that resembles typical metal-poor halo stars, including its carbon signature, and no exceptional over- or underabundances. In contrast, other element ratios, $[X/Fe]$, are somewhat enhanced in HE 1327–2327 with respect to stars with $-4.0 < [Fe/H] < -2.5$ but less so for the giants HE 0107–5240 and HE 0557–4840. No neutron-capture element was detected in HE 0107–5240, HE 0557–4840, or SDSS J102915+172927, whereas, unexpectedly, a large value of $[Sr/Fe] = 1.1$ was obtained for HE 1327–2326. Despite expectations, and as discussed in [Sect. 4.1.2](#), lithium was not detected in either the relatively unevolved subgiant HE 1327–2326 or the dwarf SDSS J102915+172927. The lithium abundance upper limits are $\log \epsilon(Li) < 0.7$ (Frebel et al. 2008) and < 1.1 (Caffau et al. 2011), respectively. These results are extremely surprising. Given that HE 1327–2326 and SDSS J102915+172927 have $T_{\text{eff}} = 6,180$ K and 5,810 K, respectively, one would expect them to lie on the Spite plateau, with $\log \epsilon(Li) = 2.3$. Somewhat unsatisfactory conjectures that might explain the non-detection include (1) Li at the epoch of lowest metallicity was below the abundance of the Spite Plateau due to its destruction early in the Universe (see e.g., Piau et al. 2006 for an interesting scenario) and (2) Li was destroyed by phenomena associated with (not yet detected) binarity. Progress will probably only be made when more near-main-sequence-turnoff stars with $[Fe/H] \lesssim -4.0$ are discovered which permit clarification of this issue.

Both HE 0107–5240 and HE 1327–2326 are benchmark objects with the potential to constrain various theoretical studies of the early Universe, such as the formation of the first stars, calculations of Population III SN yields, and the earliest chemical evolution. Several different scenarios have been offered that seek to explain the highly individual abundance patterns of both stars as early, extreme Population II, stars that display the “fingerprint” of just one Population III SN. These include (1) “mixing and fallback” models (Umeda and Nomoto 2003; Iwamoto et al. 2005; Nomoto et al. 2006) of a faint (i.e., low energy) $25 M_{\odot}$ supernova in which a large amount of C, N, and O but little Fe is ejected, while a large fraction of Fe-rich ejecta is postulated to fall back onto the newly created black hole. (See [Fig. 3-22](#) for comparison of the observed and predicted abundances for HE 0107–5240 and HE 1327–2326); (2) the modeling of Heger and Woosley (2010) who fit the observed stellar abundances by searching for a match within a large grid of Population III SN yields. Their best fit involved typical halo stars with a power-law IMF in the range $M = 11–15 M_{\odot}$, low explosion energy, and little mixing; and (3) the investigation of Meynet et al. (2006) who explored the influence of stellar rotation on elemental yields of $60 M_{\odot}$ near-zero-metallicity SNe. Mass loss from rotating massive Population III stars qualitatively reproduces the CNO abundances observed in HE 1327–2326 and other carbon-rich metal-poor stars. In a somewhat different model, Suda et al. (2004) proposed a scenario in which the abundances of HE 0107–5240 originated in a Population III binary system that experienced mass transfer of CNO elements from the more massive companion during its AGB phase, together with subsequent accretion of heavy elements from the ISM onto the (less massive) component now being observed. Along the same lines, Campbell et al. (2010) suggested a



■ Fig. 3-22

Abundance distribution versus atomic number for the two most Fe-poor stars HE 0107–5240 and HE 1327–2326 (circle and triangles, respectively) compared with the best fit models of “mixing and fallback” core-collapse SNe (from Nomoto et al. 2006). The middle dotted line shows the solar abundance ratio. See text for more details on the SNe models

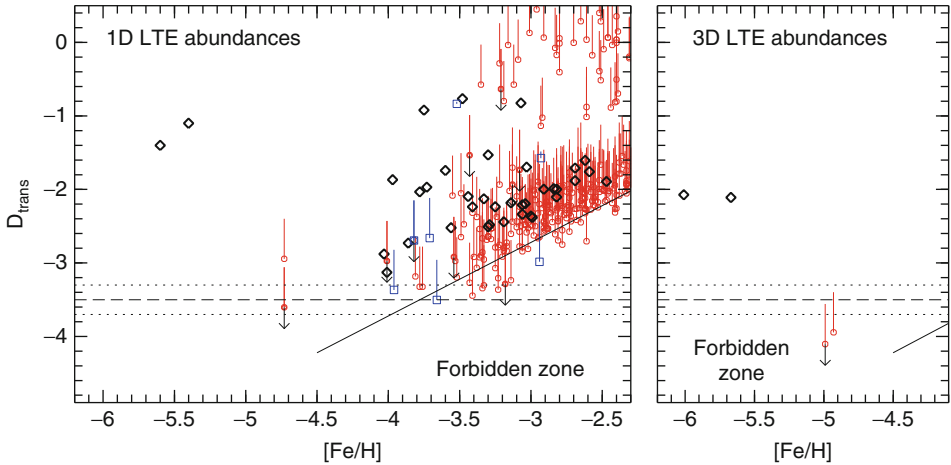
binary model for HE 1327–2326 in terms of s-process nucleosynthesis and mass transfer via a stellar wind. Qualitatively, the high C, N, O, and Sr stellar abundances could be explained this way if the star was in a wide binary.

Stars with $[\text{Fe}/\text{H}] > -4.3$ and “classical” halo abundance signatures have also been reproduced with Population III SN yields. Average abundance patterns of four non-carbon-enriched stars with $-4.2 < [\text{Fe}/\text{H}] < -3.5$ were modeled with the yields of massive ($\sim 30\text{--}50 M_{\odot}$), high explosion energy ($\sim 20\text{--}40 \times 10^{51}$ ergs), Population III hypernovae (Tominaga et al. 2007), and also fit with integrated yields of a small number of Population III stars (Heger and Woosley 2010). Special types of SNe or unusual nucleosynthesis yields have been considered for stars with chemically peculiar abundances, for example, high Mg. It is, however, often difficult to explain the entire abundance pattern in this way. Abundances of additional stars with $[\text{Fe}/\text{H}] < -4.0$, as well as a better understanding of the explosion mechanism and the effect of the initial conditions on SNe yields, are required to arrive at a more comprehensive picture of the extent to which metal-poor stars reflect the ejecta of the original Population III objects or, alternatively, those of later generations of SNe.

Some metal-poor stars display abundance ratios of a few elements that differ by large amounts from the general halo trend (e.g., C, Mg). The level of chemical diversity furthermore increases toward the lowest metallicities. For instance, as discussed in [Sect. 4.2.1](#), a large fraction of the most metal-poor stars is very carbon-rich (i.e., $[\text{C}/\text{Fe}] > 0.7$). In the compilation of Frebel (2010), the C-rich fraction of stars with $[\text{Fe}/\text{H}] < -2.0$ is ~ 0.17 . Most significantly, the fraction increases with decreasing metallicity (see [Fig. 3-4](#)), and indeed, three of the four stars with $[\text{Fe}/\text{H}] < -4.3$ are extremely carbon-rich. Reasons for this general behavior remain unclear. Could there be a cosmological origin for the large fraction of carbon-rich stars? Ideas for the required cooling processes necessary to induce sufficient fragmentation of the near-primordial gas to enable low-mass star formation include cooling based on enhanced molecule formation due to ionization of the gas, cooling through metal enrichment or dust, and complex effects such as turbulence and magnetic fields (Bromm et al. 2009). Fine-structure line cooling through C II and O I was suggested as a main cooling agent (Bromm and Loeb 2003). These elements were likely produced in vast quantities in Population III objects (see [Sect. 4.2.1](#)) and may have been responsible for the ISM reaching a critical metallicity, sufficient for low-mass star formation.

The existence and level of such a “critical metallicity” can be probed with large numbers of carbon and oxygen-poor metal-poor stars: if a threshold exists, all of these objects should have a combination of C and/or O abundances *above* the threshold for a critical metallicity. A transition discriminant was defined by Frebel et al. (2007b), which has since been slightly revised to $D_{\text{trans}} = \log(10^{[\text{C}/\text{H}]} + 0.9 \times 10^{[\text{O}/\text{H}]})$ (V. Bromm private communication). No low-mass metal-poor stars should exist below the critical value of $D_{\text{trans}} = -3.5$.

As can be seen in [Fig. 3-23](#), at metallicities of $[\text{Fe}/\text{H}] \gtrsim -3.5$, most stars have C and/or O abundances that place them well above the threshold. They simply follow the solar C and O abundances scaled down to their respective Fe values. Clearly, this metallicity range is not suitable for directly probing the very early time. Below $[\text{Fe}/\text{H}] \sim -3.5$, however, the observed C and/or O levels must be *higher* than the Fe-scaled solar abundances to be above the critical metallicity. Indeed, apart from one object, none of the known lowest-metallicity stars appear to have D_{trans} values or limits below the critical value, consistent with this cooling theory. The exception is SDSS J102915+172927, which has an upper limit for carbon of only $[\text{C}/\text{H}] < -3.8$ (1D) and < -4.3 (3D). Assuming the above $[\text{C}/\text{O}]$ range, this leads to $D_{\text{trans}} < (-3.6 \text{ to } -3.0)$




■ Fig. 3-23

Left panel: Transition discriminant, D_{trans} , for Galactic halo (small red circles, thick black diamonds) and dwarf Galaxy (blue squares) metal-poor stars as a function of $[\text{Fe}/\text{H}]$, based on 1D abundances. Black diamonds show stars with D_{trans} values calculated from their C and O abundances. Red circles and blue squares depict lower D_{trans} limits based on only a known C abundance. The corresponding vertical bars show the potential range of D_{trans} for a given star assuming O to be tied to C within the range $-0.7 < [\text{C}/\text{O}] < 0.2$. (If an upper limit on O is available and less than the maximal assumed O abundance, the bar is correspondingly shorter.) Circles or squares with bars plus additional arrows indicate interesting cases where only upper limits of C abundances are available and nothing is known about the O abundance. The solid line represents the solar C and O abundances scaled down with $[\text{Fe}/\text{H}]$, while dashed and dotted lines display the transition discriminant $D_{\text{trans}} = -3.5$ together with uncertainties. The “Forbidden zone” indicates the region with insufficient amounts of C and O for low-mass star formation (Based on Fig. 1 of Frebel et al. 2007b with recent additions from the literature such as Caffau et al. 2011). **Right panel:** Same as left panel, but for $[\text{Fe}/\text{H}] < -4.1$ and using 3D carbon and oxygen abundances, for the four most iron-poor stars

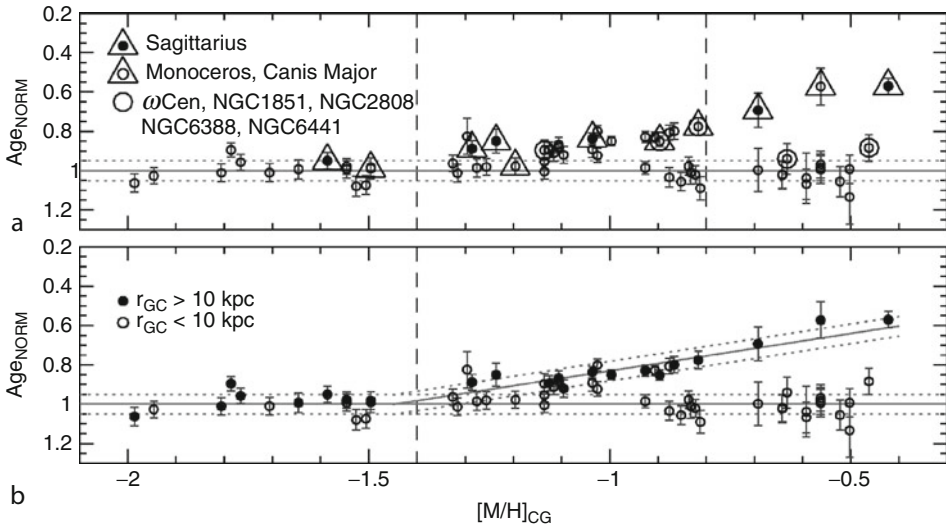
(1D) and $< (-4.1 \text{ to } -3.5)$ (3D). However, only with a known O abundance can this low D_{trans} value be conclusively determined. As can be seen in ► Fig. 3-23, several other stars also have values that are close to $D_{\text{trans}} = -3.5$, based on their 1D abundances. Very high S/N spectra suitable for measurements of very weak CH and OH molecular features will be required to determine exactly how close the D_{trans} of these objects are to the critical value.


The likely exception of SDSS J102915+172927 and several interesting “border line” cases notwithstanding, this cooling theory suggests that at low metallicity, carbon excesses are a requirement for the formation of most low-mass metal-poor stars. This is qualitatively in line with the empirical finding of a large fraction of carbon-rich stars and may thus reflect a generic avenue for low-mass star formation. Individual objects, of course, could be the result of unusual circumstances or different mechanism. For example, SDSS J102915+172927 could have formed from a gas cloud that was primarily cooled by dust grains, rather than atoms, made by the first stars. If future data show that the majority of the most metal-poor stars have $D_{\text{trans}} < -3.5$, then dust cooling (inducing a much lower critical metallicity) would be a dominant cooling mechanism in the early Universe.

6.2 The Milky Way

During the past half century, two basically different observationally driven paradigms were proposed for the formation of the Galactic halo. The first was the monolithic collapse model of Eggen et al. (1962) (hereafter ELS) and the second the accretion model of Searle and Zinn (1978) (hereafter SZ). At the same time, White and Rees (1978) proposed, in a more general context, their CDM hierarchical clustering paradigm in which “The entire luminosity content of galaxies ... results from the cooling and fragmentation of residual gas within the transient potential wells provided by the dark matter.” ELS predicted a very rapid collapse phase (of a few 10^8 year) and a dependence of kinematics on abundance together with a radial abundance gradient for halo material. SZ, in contradistinction, predicted a longer formation period of a few 10^9 years, no dependence of kinematics on abundance, and no radial abundance gradient. Not too surprisingly, perhaps, neither gives a complete explanation of the more complicated reality. On the one hand, early work revealed no dependence of kinematics on abundance for $[\text{Fe}/\text{H}] \lesssim -1.7$ (see Chiba and Beers 2000, and references therein), while on the other, globular cluster age measurements demonstrated that although some clusters were significantly younger than the majority, the age spread was small for the bulk of the system.  Figure 3-24, from the recent work of Marín-Franch et al. (2009), presents the relative ages of 64 clusters as a function of metallicity, $[\text{M}/\text{H}]$, which illustrates this point.

A turning point in the discussion came with the discovery by Ibata et al. (1995) of the Sagittarius dwarf galaxy, which has been captured by the Milky Way and is currently being torn apart in its gravitational field. Some six of the Galactic globular clusters are believed to have once been part of the Sgr system. Marín-Franch et al. (2009) comment on similar over-densities in Monoceros and Canis Major that may contain several other globular clusters and be associated



 Fig. 3-24

Relative ages, Age_{NORM} , for the Galactic globular clusters as a function of cluster metallicity, $[\text{M}/\text{H}]$ (where $[\text{M}/\text{H}] = [\text{Fe}/\text{H}] + \log(0.638f + 0.362)$, and $\log(f) = [\alpha/\text{Fe}]$), from the work of Marín-Franch et al. (2009, Fig. 13). Approximate absolute ages may be obtained as $12.8 \times \text{Age}_{\text{NORM}}$

with similar accretions. Against this background, it is then very instructive to consider the detail of [Fig. 3-24](#). [Marín-Franch et al. \(2009\)](#) identify two groups of globular clusters: “a population of old clusters with an age dispersion of $\sim 5\%$ (i.e., ~ 0.6 Gyr) and no age-metallicity relationship, and a group of younger clusters with an age-metallicity relationship similar to that of the globular clusters associated with the Sagittarius dwarf galaxy.” Two thirds of the sample belong to the old group, one third to the younger.

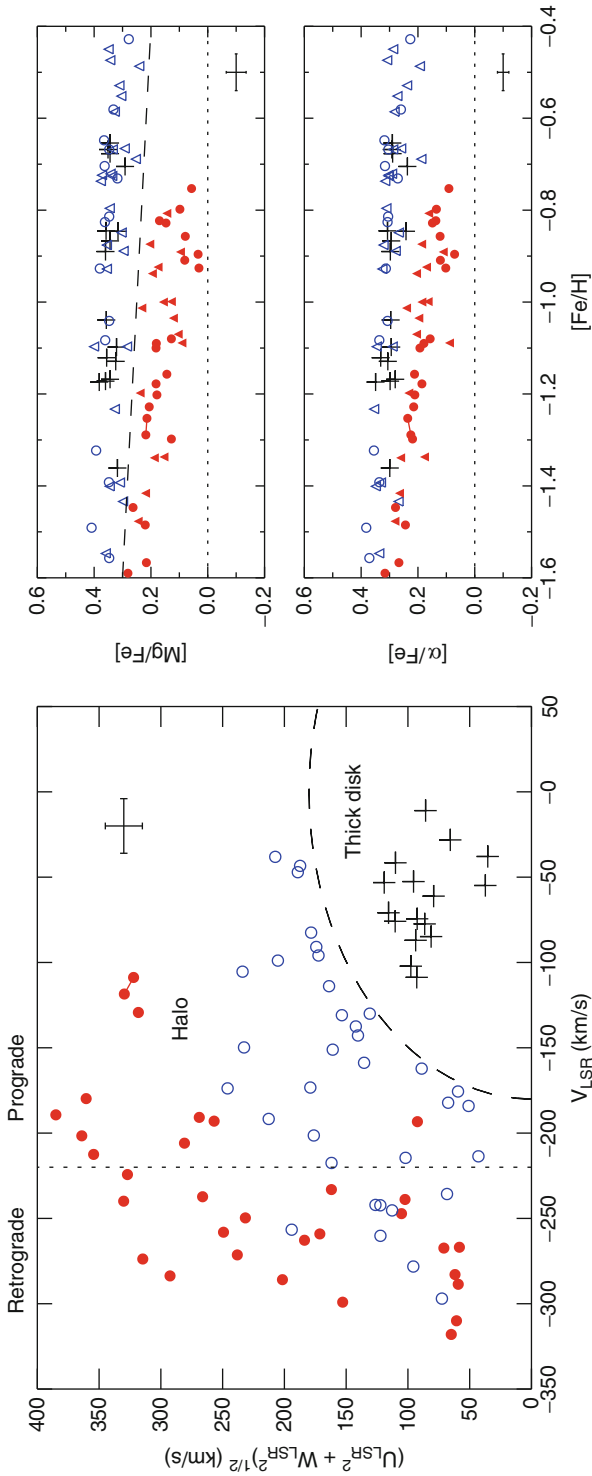
As noted above in [Sect. 3.2.1](#), there has been growing evidence that the field halo stars of the Milky Way comprise more than one population. The reader should consult [Carollo et al. \(2010\)](#) for the development of the case that the Galaxy’s halo contains an inner and an outer component. They report the following essential differences between the two components, which are dominant interior and exterior to ~ 15 kpc: (1) the inner component is more flattened than the outer component, with axial ratio (c/a) values of ~ 0.6 and 1.0 , respectively; (2) the inner component has small prograde systemic rotation, $\langle V_\phi \rangle = +7 \pm 4 \text{ km s}^{-1}$ (i.e., rotating in the same sense as the Galactic disk), while the outer has retrograde rotation $\langle V_\phi \rangle = -80 \pm 13 \text{ km s}^{-1}$; and (3) the inner component is more metal-rich, with peak metallicity $[\text{Fe}/\text{H}] = -1.6$, while the outer one has a peak metallicity $[\text{Fe}/\text{H}] = -2.2$.

Against this background, [Morrison et al. \(2009\)](#) have reported another, more highly flattened halo component, with $c/a \sim 0.2$, which has “a small prograde rotation ... supported by velocity anisotropy, and contains more intermediate-metallicity stars (with $-1.5 < [\text{Fe}/\text{H}] < -1.0$) that the rest of [the] sample.”

While the detailed nature and relationships of these components remain to be fully understood, it seems likely the answer will be found within the hierarchical Λ CDM paradigm reported above. The work of [Zolotov et al. \(2009\)](#), for example, while supporting the SZ paradigm of halo formation, also produces a dual halo configuration of “*in situ*” and “*accreted*” components, not unlike those envisaged in the ELS and SZ observational paradigms. Remarkably, these paradigms were first established on essentially observational grounds only. They are now being explained in terms of a theoretical framework based on tracing the dark-matter evolution from initial density fluctuations early in the Universe.

Further support for a two component model comes from recent work of [Nissen and Schuster \(2010\)](#), who have investigated the abundances of α -elements in the abundance range $-1.6 < [\text{Fe}/\text{H}] < -0.4$, as a function of kinematics, with a view to comparing (in [Zolotov et al. 2009](#) terminology) the “*in situ*” and “*accreted*” components. Their results are shown in [Fig. 3-25](#). In the right panel, one sees a large spread in $[\alpha/\text{Fe}]$, at fixed $[\text{Fe}/\text{H}]$, that correlates strongly with position in the kinematic (so-called Toomre) diagram on the left. The simplest and also extremely significant interpretation of this figure is that stars with $[\alpha/\text{Fe}] \sim +0.3$ to $+0.4$, with prograde kinematics, are part of the “*in situ*” component, while those with $[\alpha/\text{Fe}] \lesssim +0.3$, on principally retrograde orbits, belong to the “*accreted*” component. The reader will recall from [Sect. 4.3.2](#) that low $[\alpha/\text{Fe}]$ is a key signature of the Milky Way’s dwarf galaxies in the range $-1.5 < [\text{Fe}/\text{H}] < 0.0$ (see also [Tolstoy et al. 2009](#), their Fig. 11). Said differently, [Fig. 3-25](#) is consistent with the view that dwarf galaxies have played an important role in the formation of the Milky Way halo.

A complementary way to study the origin of the Milky Way, its halo, and similar large galaxies more generally is through large-scale Λ CDM simulation of the growth of structure formation. A prominent issue with Milky Way-size halos at redshift $z = 0$ is the predicted large number of CDM substructures that surround such a galaxy. The number of observed dwarf galaxies surrounding the Milky Way is, however, much lower and does not agree with such predictions. This mismatch has been termed the “missing-satellite” problem (e.g., [Moore et al. 1999](#)).



■ Fig. 3-25

Left: Kinematics (in which U, V, W are velocity components in the Galactic frame). Right: $[\alpha/\text{Fe}]$ versus $[\text{Fe}/\text{H}]$ for metal-poor Milky Way stars, from the work of Nissen and Schuster (2010). Circles and triangles refer to halo (and crosses to thick disk) stars, respectively. Note the strongly correlated relative positions of the filled and open circles in the two panels. See text for discussion

It is thus apparent that many more questions about galaxy assembly and evolution still need to be resolved. Crucially, it remains to be seen to what extent small dark halos contained baryonic matter, subsequently observed as gas and stars, and how they evolved with time. One way to learn about the luminous content of small sub-halos is to investigate in detail the surviving dwarf galaxies, in particular the ultra-faint systems, that orbit the Milky Way. Studying the onset of star formation and associated chemical evolution in these satellites will provide some of the currently missing information for our understanding of how the observed properties of small, faint systems relate to the dark-matter substructures that built up larger galaxies.

7 Conclusions and Future Prospects

Old metal-poor stars can be employed as tools to learn about the conditions in the early Universe. The scientific topics that can be addressed in this way are numerous, and this chapter describes the most prominent questions to which metal-poor stars can provide unique insights. These include the origin and evolution of the chemical elements, the relevant nucleosynthesis processes and sites, and the overall chemical and dynamical history of the Galaxy. By extension, the abundance patterns in metal-poor stars provide constraints on the nature of the first stars and the initial mass function, and the chemical yields of first/early SNe. Moreover, studying metal-poor stars in dwarf galaxies open up ways to learn about early star and early galaxy formation processes, including the formation of the Galactic halo through hierarchical assembly.

Our review has highlighted the tension between the approximations inherent in 1D/LTE model atmosphere abundance analyses, on the one hand, and the more physically realistic 3D/non-LTE (and more computationally challenging) formalism, on the other. Given abundance differences $\sim 0.5\text{--}0.9$ dex between the two formalisms for many elements, there is an urgent need for comprehensive investment in self-consistent 3D/non-LTE modeling of the relevant regions of $T_{\text{eff}}/\log g/[\text{Fe}/\text{H}]$ space.

The most metal-deficient stars are extremely rare, but past surveys for metal-poor halo stars have shown that they can be systematically identified through several selection steps. Typically, the metal-poor halo stars found to date are located no further away than $\sim 10\text{--}15$ kpc, with $B \lesssim 16$. This brightness limit ensures that adequate S/N spectra can be obtained in reasonable observing times with existing telescope/instrument combinations. The outer halo beyond ~ 15 kpc, however, is largely unexplored territory, at least in terms of high-resolution spectroscopy. Recent work has shown that the ultra-faint dwarf galaxies orbiting the Galaxy contain larger fractions of extremely metal-poor stars (i.e., $[\text{Fe}/\text{H}] < -3.0$) than does the Galactic halo. That said, there is a high price to be paid in order to observe these objects because most of them are extremely faint. Those that are currently observable with 6–10-m telescopes are only the brightest in a given system, and these are usually located on the upper RGB. At the limit are objects at 19th magnitude that can just be observed at high spectral resolution, requiring exposure times up to ~ 10 h per star in order to reach the minimum useful S/N ratio in the final spectrum. This is feasible only for individual stars, not for large-scale investigations. Objects lower on the RGB or even the main sequence (>21 mag) are out of reach, even for medium-resolution studies.

Over the next few years, all of these brightest dwarf galaxy stars will have been observed. What then? Either new larger telescopes, or additional dwarf galaxies that harbor more observable stars, are required. Addressing both options is currently underway. To chemically characterize the Galactic halo in detail (including its streams, substructures, and satellites) wide-angle surveys with large volumes are needed. The Australian SkyMapper photometric survey (which began surveying in 2011) is optimized for stellar work. It will provide a wealth of metal-poor candidates in need of detailed high-resolution follow-up to determine their abundances. The footprint of this project will be some three times larger than that of the HES. Newly discovered stars with $B \lesssim 16$ will enable an important advance in stellar archaeology by (hopefully) trebling the number of “bright” objects available for high-resolution abundance studies with existing facilities. A significant fraction of SkyMapper candidates will, however, be too faint for practical and efficient follow-up observations. In particular, the most metal-poor stars require high S/N to enable the detection of very weak absorption features. These will be the target of the high-resolution spectrographs on the new generation of 20–40-m telescopes. Among these new discoveries, it is expected that more of the most metal-poor stars (e.g., those with $[Fe/H] < -5.0$; Christlieb et al. 2002 and Frebel et al. 2005) will be found.

With SkyMapper, also many more faint dwarf galaxies are expected to be found. Even though the brightest stars in them will still be at the observational limit for high-resolution spectroscopy, having more of these dwarf galaxy stars available for detailed studies will provide new insights into the nature and evolution of these small systems and their relationship to the building-up process(es) of the Milky Way. Other photometric surveys such as Pan-Starrs and LSST are also expected to yield new dwarf galaxies. These surveys, however, will be useful for the search for metal-poor stars in dwarf galaxies only if coupled with additional follow-up efforts, due to the lack of sufficiently metal-sensitive filters.

In addition to these photometric surveys, the Chinese LAMOST spectroscopic survey will provide numerous metal-poor candidates in the northern hemisphere, all based on medium-resolution spectra. GAIA is an astrometric space mission led by ESA, scheduled to begin observations in 2013. It will obtain high-precision phase-space information for one billion stars in the Galaxy, along with the physical parameters and the chemical composition of many of these stars. These new data will revolutionize our understanding of the origin, evolution, structure, and dynamics of the Milky Way as a whole and of its components. In particular, the kinematic information (e.g., proper motions) that will become available for many known metal-poor stars will enable detailed studies of how the abundances of different populations depend on kinematics. Furthermore, a precise selection of low-metallicity candidate stars based on, for example, extreme kinematic signatures, will become feasible. Since this is currently beyond reach for most metal-poor halo giants, the GAIA astrometry should increase the yield of fainter metal-poor stars at larger distances.

By having the opportunity to access fainter stars in the outer Galactic halo and dwarf galaxies, the next major frontier in stellar archaeology and near-field cosmology can be tackled. High-resolution follow-up of faint stars may become a reality with the light-collecting power of the next generation of optical telescopes: the Giant Magellan Telescope, the Thirty Meter Telescope, and the European Extremely Large Telescope. All three telescopes are currently in the planning and design phase with completions scheduled around 2020. It is currently envisaged that the Giant Magellan Telescope will be equipped with a high-resolution spectrograph at first light. This facility would then not only permit in-depth analysis of new metal-poor stars in the Galaxy’s outer halo and dwarf galaxies but also make it feasible to obtain very high- S/N

data of somewhat brighter stars; to permit investigation, for example, of isotopic ratios such as ${}^6\text{Li}/{}^7\text{Li}$ and r-process-enhanced stars; and to provide crucial empirical constraints on the nature of the site and details of critical nucleosynthesis processes. This is currently possible only for the very brightest stars. At the faintest magnitudes, individual stars in the Magellanic Clouds will become accessible for high-resolution spectroscopy. In the event of such a spectrograph being available in the northern hemisphere, then perhaps even the brightest objects in Andromeda could be observed. Studying massive dwarf galaxies and another spiral system that resembles the Milky Way would provide unprecedented new insight into the chemical evolution of large systems and their formation process(es).

All of these new observations will be accompanied by an increased theoretical understanding of the first stars and galaxies, SN nucleosynthesis, the mixing of metals into the existing gaseous medium, and feedback effects in the early Universe, as well as cosmic chemical evolution. New generations of hydrodynamical, high-resolution cosmological simulations will enable a direct investigation of chemical evolution by including more than one SN and corresponding feedback(s), for example, in first-galaxy simulations. These will be sufficient for tracing the corresponding metal production and spatial distributions and enable direct comparisons with abundance measurements in the stars of dwarf galaxies. This in turn will shed new light on the question of whether the least-luminous dwarf galaxies resemble the first galaxies and if they are early analogs of the building blocks of the Galactic halo.

Acknowledgements

A.F. acknowledges support from a Clay Fellowship (administered by the Smithsonian Astrophysical Observatory) and the Harvard-Smithsonian Center for Astrophysics for enabling her to prepare this chapter. It is a pleasure to thank M. Asplund, T. C. Beers, N. Christlieb, G. L. Harris, A. Karakas, and H. L. Morrison for their perceptive comments, which led to improvement to the manuscript. The authors also thank N. Christlieb and E. N. Kirby for supplying the high-resolution spectrum of HE 0107–5240 presented in [▶ Figs. 3-2](#) and [▶ 3-10](#), respectively.

Cross-References

- [▶ Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- [▶ Dynamics of Disks and Warps](#)
- [▶ Galactic Distance Scales](#)
- [▶ Globular Cluster Dynamical Evolution](#)
- [▶ High-Velocity Clouds](#)
- [▶ Interstellar PAHs and Dust](#)
- [▶ Mass Distribution and Rotation Curve in the Galaxy](#)
- [▶ Open Clusters and their Role in the Galaxy](#)
- [▶ Star Counts and the Nature of the Galactic Thick Disk](#)
- [▶ The Galactic Bulge](#)
- [▶ The Stellar and Sub-Stellar Initial Mass Function of Simple and Composite Populations](#)

References

- Aoki, W., Arimoto, N., Sadakane, K., Tolstoy, E., Battaglia, G., Jablonka, P., Shetrone, M., Letarte, B., Irwin, M., Hill, V., Francois, P., Venn, K., Primas, F., Helmi, A., Kaufer, A., Tafelmeyer, M., Szeifert, T., & Babusiaux, C. 2009, *A&A*, 502, 569
- Aoki, W., Beers, T. C., Christlieb, N., Norris, J. E., Ryan, S. G., & Tsangarides, S. 2007, *ApJ*, 655, 492
- Aoki, W., Norris, J. E., Ryan, S. G., Beers, T. C., & Ando, H. 2002, *ApJ*, 576, L141
- Aoki, W., Ryan, S. G., Norris, J. E., Beers, T. C., Ando, H., Iwamoto, N., Kajino, T., Mathews, G. J., & Fujimoto, M. Y. 2001, *ApJ*, 561, 346
- Arlandini, C., Käppeler, F., Wisshak, K., Gallino, R., Lugaro, M., Busso, M., & Straniero, O. 1999, *ApJ*, 525, 886
- Arnett, D. 1996, *Supernovae and Nucleosynthesis. An Investigation of the History of Matter, from the Big Bang to the Present* (Princeton: Princeton University Press)
- Asplund, M. 2005, *ARA&A*, 43, 481
- Asplund, M., Carlsson, M., & Botnen, A. V. 2003, *A&A*, 399, L31
- Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, *ARA&A*, 47, 481
- Baade, W. 1944, *ApJ*, 100, 137
- Bahcall, J. N., & Soneira, R. M. 1980, *ApJS*, 44, 73
- Barklem, P. S., Christlieb, N., Beers, T. C., Hill, V., Bessell, M. S., Holmberg, J., Marsteller, B., Rossi, S., Zickgraf, F.-J., & Reimers, D. 2005, *A&A*, 439, 129
- Beers, T. C., & Christlieb, N. 2005, *ARA&A*, 43, 531
- Beers, T. C., Preston, G. W., & Shectman, S. A. 1992, *AJ*, 103, 1987
- Behr, B. B., Cohen, J. G., McCarthy, J. K., & Djorgovski, S. G. 1999, *ApJ*, 517, L135
- Bergemann, M., Pickering, J. C., & Gehren, T. 2010, *MNRAS*, 401, 1334
- Bessell, M. S., & Norris, J. 1984, *ApJ*, 285, 622
- Bromm, V., & Larson, R. B. 2004, *ARA&A*, 42, 79
- Bromm, V., & Loeb, A. 2003, *Nature*, 425, 812
- Bromm, V., Yoshida, N., Hernquist, L., & McKee, C. F. 2009, *Nature*, 459, 49
- Burbidge, E. M., Burbidge, G. R., Fowler, W. A., & Hoyle, F. 1957, *Rev. Modern Phys.*, 29, 547
- Burris, D. L., Pilachowski, C. A., Armandroff, T. E., Snedden, C., Cowan, J. J., & Roe, H. 2000, *ApJ*, 544, 302
- Caffau, E., Bonifacio, P., François, P., Sbordone, L., Monaco, L., Spite, M., Spite, F., Ludwig, H. -G., Cayrel, R., Zaggia, S., Hammer, F., Randich, S., Molaro, P., & Hill, V. 2011, *Nature*, 477, 67
- Campbell, S. W., Lugaro, M., & Karakas, A. I. 2010, *A&A*, 522, L6
- Carney, B. W., Laird, J. B., Latham, D. W., & Aguilar, L. A. 1996, *AJ*, 112, 668
- Carney, B. W., & Peterson, R. C. 1981, *ApJ*, 245, 238
- Carollo, D., Beers, T. C., Chiba, M., Norris, J. E., Freeman, K. C., Lee, Y. S., Ivezić, Ž., Rockosi, C. M., & Yanny, B. 2010, *ApJ*, 712, 692
- Carretta, E., Bragaglia, A., Gratton, R., D'Orazi, V., & Lucatello, S. 2009, *A&A*, 508, 695
- Cayrel, R., Depagne, E., Spite, M., Hill, V., Spite, F., François, P., Plez, B., Beers, T., Primas, F., Andersen, J., Barbuy, B., Bonifacio, P., Molaro, P., & Nordström, B. 2004, *A&A*, 416, 1117
- Chamberlain, J. W., & Aller, L. H. 1951, *ApJ*, 114, 52
- Chiba, M., & Beers, T. C. 2000, *AJ*, 119, 2843
- Christlieb, N., Bessell, M. S., Beers, T. C., Gustafsson, B., Korn, A., Barklem, P. S., Carlsson, T., Mizuno-Wiedner, M., & Rossi, S. 2002, *Nature*, 419, 904
- Christlieb, N., Schörck, T., Frebel, A., Beers, T. C., Wisotzki, L., & Reimers, D. 2008, *A&A*, 484, 721
- Cohen, J. G., & Huang, W. 2009, *ApJ*, 701, 1053
- Collet, R., Asplund, M., & Trampedach, R. 2006, *ApJ*, 644, L121
- Conroy, C., & Spergel, D. N. 2011, *ApJ*, 726, 36
- Cooke, R., Pettini, M., Steidel, C. C., Rudie, G. C., & Jorgenson, R. A. 2011, *MNRAS*, 412, 1047
- Cybert, R. H., Fields, B. D., & Olive, K. A. 2008, *J. Cosmol. Astropart. Phys.*, 11, 12
- Diemand, J., Kuhlen, M., & Madau, P. 2007, *ApJ*, 667, 859
- Dupree, A. K., Strader, J., & Smith, G. H. 2011, *ApJ*, 728, 155
- Eggen, O. J., Lynden-Bell, D., & Sandage, A. R. 1962, *ApJ*, 136, 748
- Fabbian, D., Asplund, M., Barklem, P. S., Carlsson, M., & Kiselman, D. 2009, *A&A*, 500, 1221
- François, P., Depagne, E., Hill, V., Spite, M., Spite, F., Plez, B., Beers, T. C., Andersen, J., James, G., Barbuy, B., Cayrel, R., Bonifacio, P., Molaro, P., Nordström, B., & Primas, F. 2007, *A&A*, 476, 935
- Frebel, A. 2010, *Astron. Nachr.*, 331, 474
- Frebel, A., Aoki, W., Christlieb, N., Ando, H., Asplund, M., Barklem, P. S., Beers, T. C., Eriksson, K., Fechner, C., Fujimoto, M. Y., Honda, S., Kajino, T., Minezaki, T., Nomoto, K., Norris, J. E., Ryan, S. G., Takada-Hidai, M., Tsangarides, S., & Yoshii, Y. 2005, *Nature*, 434, 871
- Frebel, A., Christlieb, N., Norris, J. E., Beers, T. C., Bessell, M. S., Rhee, J., Fechner, C., Marsteller, B., Rossi, S., Thom, C., Wisotzki, L., & Reimers, D. 2006, *ApJ*, 652, 1585
- Frebel, A., Christlieb, N., Norris, J. E., Thom, C., Beers, T. C., & Rhee, J. 2007a, *ApJ*, 660, L117

- Frebel, A., Collet, R., Eriksson, K., Christlieb, N., & Aoki, W. 2008, *ApJ*, 684, 588
- Frebel, A., Johnson, J. L., & Bromm, V. 2007b, *MNRAS*, 380, L40
- Frebel, A., Johnson, J. L., & Bromm, V. 2009, *MNRAS*, 392, L50
- Frebel, A., Kirby, E. N., & Simon, J. D. 2010a, *Nature*, 464, 72
- Frebel, A., Simon, J. D., Geha, M., & Willman, B. 2010b, *ApJ*, 708, 560
- Freeman, K., & Bland-Hawthorn, J. 2002, *ARA&A*, 40, 487
- Fulbright, J. P., Rich, R. M., & Castro, S. 2004, *ApJ*, 612, 447
- Geisler, D., Smith, V. V., Wallerstein, G., Gonzalez, G., & Charbonnel, C. 2005, *AJ*, 129, 1428
- Gilmore, G., Wyse, R. F. G., & Kuijken, K. 1989, *ARA&A*, 27, 555
- Giridhar, S., Lambert, D. L., Reddy, B. E., Gonzalez, G., & Yong, D. 2005, *ApJ*, 627, 432
- Gratton, R. G., Sneden, C., Carretta, E., & Bragaglia, A. 2000, *A&A*, 354, 169
- Gray, D. F. 2005, *The Observation and Analysis of Stellar Photospheres* (3rd ed.; Cambridge: Cambridge University Press)
- Gustafsson, B., Edvardsson, B., Eriksson, K., Jørgensen, U. G., Nordlund, Å., & Plez, B. 2008, *A&A*, 486, 951
- Hartwick, F. D. A. 1976, *ApJ*, 209, 418
- Hartwick, F. D. A. 1987, in *NATO ASIC Proc. 207: The Galaxy*, ed. G. Gilmore, & B. Carswell (Dordrecht/Boston: D. Reidel), 281
- Heger, A., & Woosley, S. E. 2002, *ApJ*, 567, 532
- Heger, A., & Woosley, S. E. 2010, *ApJ*, 724, 341
- Hill, V., Plez, B., Cayrel, R., Nordström, T. B. B., Andersen, J., Spite, M., Spite, F., Barbuy, B., Bonifacio, P., Depagne, E., François, P., & Primas, F. 2002, *A&A*, 387, 560
- Honda, S., Aoki, W., Ishimaru, Y., Wanajo, S., & Ryan, S. G. 2006, *ApJ*, 643, 1180
- Honda, S., Aoki, W., Kajino, T., Ando, H., Beers, T. C., Izumiura, H., Sadakane, K., & Takada-Hidai, M. 2004, *ApJ*, 607, 474
- Ibata, R. A., Gilmore, G., & Irwin, M. J. 1995, *MNRAS*, 277, 781
- Iwamoto, N., Umeda, H., Tominaga, N., Nomoto, K., & Maeda, K. 2005, *Science*, 309, 451
- Jonsell, K., Barklem, P. S., Gustafsson, B., Christlieb, N., Hill, V., Beers, T. C., & Holmberg, J. 2006, *A&A*, 451, 651
- Karakas, A., & Lattanzio, J. C. 2007, *PASA*, 24, 103
- Kirby, E. N., Cohen, J. G., Smith, G. H., Majewski, S. R., Sohn, S. T., & Guhathakurta, P. 2011, *ApJ*, 727, 79
- Kirby, E. N., Simon, J. D., Geha, M., Guhathakurta, P., & Frebel, A. 2008, *ApJ*, 685, L43
- Kobayashi, C., Umeda, H., Nomoto, K., Tominaga, N., & Ohkubo, T. 2006, *ApJ*, 653, 1145
- Koch, A., McWilliam, A., Grebel, E. K., Zucker, D. B., & Belokurov, V. 2008, *ApJ*, 688, L13
- Korn, A. J., Grundahl, F., Richard, O., Mashonkina, L., Barklem, P. S., Collet, R., Gustafsson, B., & Piskunov, N. 2007, *ApJ*, 671, 402
- Lai, D. K., Bolte, M., Johnson, J. A., Lucatello, S., Heger, A., & Woosley, S. E. 2008, *ApJ*, 681, 1524
- Li, H. N., Christlieb, N., Schörck, T., Norris, J. E., Bessell, M. S., Yong, D., Beers, T. C., Lee, Y. S., Frebel, A., & Zhao, G. 2010, *A&A*, 521, A10
- Lind, K., Primas, F., Charbonnel, C., Grundahl, F., & Asplund, M. 2009, *A&A*, 503, 545
- Marín-Franch, A., Aparicio, A., Piotto, G., Rosenberg, A., Chaboyer, B., Sarajedini, A., Siegel, M., Anderson, J., Bedin, L. R., Dotter, A., Hempel, M., King, I., Majewski, S., Milone, A. P., Paust, N., & Reid, I. N. 2009, *ApJ*, 694, 1498
- Martin, N. F., de Jong, J. T. A., & Rix, H. -W. 2008, *ApJ*, 684, 1075
- Mateo, M. L. 1998, *ARA&A*, 36, 435
- Meyer, B. S. 1994, *ARA&A*, 32, 153
- McWilliam, A. 1997, *ARA&A*, 35, 503
- McWilliam, A., Preston, G. W., Sneden, C., & Searle, L. 1995, *AJ*, 109, 2757
- Meléndez, J., Casagrande, L., Ramírez, I., Asplund, M., & Schuster, W. J. 2010, *A&A*, 515, L3
- Meynet, G., Ekström, S., & Maeder, A. 2006, *A&A*, 447, 623
- Moore, B., Ghigna, S., Governato, F., Lake, G., Quinn, T., Stadel, J., & Tozzi, P. 1999, *ApJ*, 524, L19
- Morrison, H. L., Helmi, A., Sun, J., Liu, P., Gu, R., Norris, J. E., Harding, P., Kinman, T. D., Kephley, A. A., Freeman, K. C., Williams, M., & Van Duyn, J. 2009, *ApJ*, 694, 130
- Neill, J. D., Sullivan, M., Gal-Yam, A., Quimby, R., Ofek, E., Wyder, T. K., Howell, D. A., Nugent, P., Seibert, M., Martin, D. C., Overzier, R., Barlow, T. A., Foster, K., Friedman, P. G., Morrissey, P., Neff, S. G., Schiminovich, D., Bianchi, L., Donas, J., Heckman, T. M., Lee, Y., Madore, B. F., Milliard, B., Rich, R. M., & Szalay, A. S. 2011, *ApJ*, 727, 15
- Nissen, P. E., & Schuster, W. J. 2010, *A&A*, 511, L10
- Nomoto, K., Tominaga, N., Umeda, H., Kobayashi, C., & Maeda, K. 2006, *Nucl. Phys. A*, 777, 424
- Norris, J. E. 1999, in *Astronomical Society of the Pacific Conference Series*, Vol. 165, *The Third Stromlo Symposium: The Galactic Halo*, ed. B. K. Gibson, R. S. Axelrod, & M. E. Putman (San Francisco: Astronomical Society of the Pacific), 213

- Norris, J. E., Ryan, S. G., & Beers, T. C. 2001, *ApJ*, 561, 1034
- Norris, J. E., Christlieb, N., Korn, A. J., Eriksson, K., Bessell, M. S., Beers, T. C., Wisotzki, L., & Reimers, D. 2007, *ApJ*, 670, 774
- Norris, J. E., Gilmore, G., Wyse, R. F. G., Yong, D., & Frebel, A. 2010a, *ApJ*, 722, L104
- Norris, J. E., Wyse, R. F. G., Gilmore, G., Yong, D., Frebel, A., Wilkinson, M. I., Belokurov, V., & Zucker, D. B. 2010b, *ApJ*, 723, 1632
- Norris, J. E., Yong, D., Gilmore, G., & Wyse, R. F. G. 2010c, *ApJ*, 711, 350
- Pagel, B. E. J. 1997, *Nucleosynthesis and Chemical Evolution of Galaxies* (Cambridge: Cambridge University Press)
- Piau, L., Beers, T. C., Balsara, D. S., Sivarani, T., Truran, J. W., & Ferguson, J. W. 2006, *ApJ*, 653, 300
- Prantzos, N. 2003, *A&A*, 404, 211
- Qian, Y.-Z., & Wasserburg, G. J. 2003, *ApJ*, 588, 1099
- Roederer, I. U., Kratz, K., Frebel, A., Christlieb, N., Pfeiffer, B., Cowan, J. J., & Sneden, C. 2009, *ApJ*, 698, 1963
- Ryan, S. G., Beers, T. C., Kajino, T., & Rosolankova, K. 2001, *ApJ*, 547, 231
- Ryan, S. G., & Norris, J. E. 1991a, *AJ*, 101, 1835
- Ryan, S. G., & Norris, J. E. 1991b, *AJ*, 101, 1865
- Ryan, S. G., Norris, J. E., & Beers, T. C. 1996, *ApJ*, 471, 254
- Ryan-Weber, E. V., Pettini, M., Madau, P., & Zych, B. J. 2009, *MNRAS*, 395, 1476
- Sandage, A. 1986, *ARA&A*, 24, 421
- Schatz, H., Toenjes, R., Pfeiffer, B., Beers, T. C., Cowan, J. J., Hill, V., & Kratz, K.-L. 2002, *ApJ*, 579, 626
- Schörck, T., Christlieb, N., Cohen, J. G., Beers, T. C., Shtetman, S., Thompson, I., McWilliam, A., Bessell, M. S., Norris, J. E., Meléndez, J., Ramírez, S., Haynes, D., Cass, P., Hartley, M., Russell, K., Watson, F., Zickgraf, F., Behnke, B., Fechner, C., Fuhrmeister, B., Barklem, P. S., Edvardsson, B., Frebel, A., Wisotzki, L., & Reimers, D. 2009, *A&A*, 507, 817
- Searle, L., & Zinn, R. 1978, *ApJ*, 225, 357
- Shetrone, M., Venn, K. A., Tolstoy, E., Primas, F., Hill, V., & Kaufer, A. 2003, *AJ*, 125, 684
- Shigeyama, T., Tsujimoto, T., & Yoshii, Y. 2003, *ApJ*, 586, L57
- Simmerer, J., Sneden, C., Cowan, J. J., Collier, J., Woolf, V. M., & Lawler, J. E. 2004, *ApJ*, 617, 1091
- Simon, J. D., Frebel, A., McWilliam, A., Kirby, E. N., & Thompson, I. B. 2010, *ApJ*, 716, 446
- Sneden, C., Cowan, J. J., & Gallino, R. 2008, *ARA&A*, 46, 241
- Sneden, C., Cowan, J. J., Lawler, J. E., Ivans, I. I., Burles, S., Beers, T. C., Primas, F., Hill, V., Truran, J. W., Fuller, G. M., Pfeiffer, B., & Kratz, K.-L. 2003, *ApJ*, 591, 936
- Sneden, C., McWilliam, A., Preston, G. W., Cowan, J. J., Burris, D. L., & Amorsky, B. J. 1996, *ApJ*, 467, 819
- Songaila, A. 2001, *ApJ*, 561, L153
- Spergel, D. N., Bean, R., Doré, O., Nolta, M. R., Bennett, C. L., Dunkley, J., Hinshaw, G., Jarosik, N., Komatsu, E., Page, L., Peiris, H. V., Verde, L., Halpern, M., Hill, R. S., Kogut, A., Limon, M., Meyer, S. S., Odegard, N., Tucker, G. S., Weiland, J. L., Wollack, E., & Wright, E. L. 2007, *ApJS*, 170, 377
- Spite, F., & Spite, M. 1982, *A&A*, 115, 357
- Spite, M., Cayrel, R., Plez, B., Hill, V., Spite, F., Depagne, E., François, P., Bonifacio, P., Barbuy, B., Beers, T., Andersen, J., Molaro, P., Nordström, B., & Primas, F. 2005, *A&A*, 430, 655
- Springel, V., Wang, J., Vogelsberger, M., Ludlow, A., Jenkins, A., Helmi, A., Navarro, J. F., Frenk, C. S., & White, S. D. M. 2008, *MNRAS*, 391, 1685
- Starkenburg, E., Hill, V., Tolstoy, E., González Hernández, J. I., Irwin, M., Helmi, A., Battaglia, G., Jablonka, P., Tafelmeyer, M., Shetrone, M., Venn, K., & de Boer, T. 2010, *A&A*, 513, A34
- Suda, T., Aikawa, M., Machida, M. N., Fujimoto, M. Y., & Iben, I. J. 2004, *ApJ*, 611, 476
- Suda, T., Katsuta, Y., Yamada, S., Suwa, T., Ishizuka, C., Komiya, Y., Sorai, K., Aikawa, M., & Fujimoto, M. Y. 2008, *PASJ*, 60, 1159
- Tegmark, M., Silk, J., Rees, M. J., Blanchard, A., Abel, T., & Palla, F. 1997, *ApJ*, 474, 1
- Timmes, F. X., Woosley, S. E., & Weaver, T. A. 1995, *ApJS*, 98, 617
- Tolstoy, E., Hill, V., & Tosi, M. 2009, *ARA&A*, 47, 371
- Tominaga, N., Umeda, H., & Nomoto, K. 2007, *ApJ*, 660, 516
- Travaglio, C., Gallino, R., Arnone, E., Cowan, J., Jordan, F., & Sneden, C. 2004, *ApJ*, 601, 864
- Tsujimoto, T., Shigeyama, T., & Yoshii, Y. 1999, *ApJ*, 519, L63
- Umeda, H., & Nomoto, K. 2003, *Nature*, 422, 871
- Umeda, H., & Nomoto, K. 2005, *ApJ*, 619, 427
- Vandenberg, D. A., Bolte, M., & Stetson, P. B. 1996, *ARA&A*, 34, 461
- Venn, K. A., Irwin, M., Shetrone, M. D., Tout, C. A., Hill, V., & Tolstoy, E. 2004, *AJ*, 128, 1177
- Wagoner, R. V., Fowler, W. A., & Hoyle, F. 1967, *ApJ*, 148, 3
- Wallerstein, G., Iben, Jr., I., Parker, P., Boesgaard, A. M., Hale, G. M., Champagne, A. E., Barnes, C. A., Käppeler, F., Smith, V. V., Hoffman, R. D., Timmes, F. X., Sneden, C., Boyd, R. N., Meyer, B. S., & Lambert, D. L. 1997, *Rev. Modern Phys.*, 69, 995

- Wanajo, S., & Ishimaru, Y. 2006, *Nucl. Phys. A*, 777, 676
- Wanajo, S., Itoh, N., Ishimaru, Y., Nozawa, S., & Beers, T. C. 2002, *ApJ*, 577, 853
- Wheeler, J. C., Sneden, C., & Truran, J. W., Jr. 1989, *ARA&A*, 27, 279
- White, S. D. M. & Rees, M. J. 1978, *MNRAS*, 183, 341
- Woodsley, S. E., & Weaver, T. A. 1995, *ApJS*, 101, 181
- Zinn, R. 1985, *ApJ*, 293, 424
- Zinn, R. 1993, in *Astronomical Society of the Pacific Conference Series*, Vol. 48, *The Globular Cluster-Galaxy Connection*, ed. G. H. Smith, & J. P. Brodie (San Francisco: Astronomical Society of the Pacific), 38
- Zolotov, A., Willman, B., Brooks, A. M., Governato, F., Brook, C. B., Hogg, D. W., Quinn, T., & Stinson, G. 2009, *ApJ*, 702, 1058

4 The Stellar and Sub-Stellar Initial Mass Function of Simple and Composite Populations

Pavel Kroupa¹ · Carsten Weidner^{2,3} · Jan Pflamm-Altenburg¹ · Ingo Thies¹ · Jörg Dabringhausen¹ · Michael Marks¹ · Thomas Maschberger^{4,5}

¹Argelander-Institut für Astronomie, Universität Bonn, Bonn, Germany

²Scottish Universities Physics Alliance (SUPA), School of Physics and Astronomy, University of St. Andrews, North Haugh, St. Andrews, UK

³Instituto de Astrofísica de Canarias, La Laguna (Tenerife), Spain

⁴Institute of Astronomy, Cambridge, UK

⁵Institut de Planétologie et d' Astrophysique de Grenoble, Grenoble Cédex 9, France

1	<i>Introduction and Historical Overview</i>	118
1.1	Solar Neighborhood	119
1.2	Star Clusters	120
1.3	Intermediate-Mass and Massive Stars	121
1.4	The Invariant IMF and Its Conflict with Theory	122
1.5	Philosophical Note	124
1.6	Hypothesis Testing	125
1.7	About This Text	126
1.8	Other IMF Reviews	127
2	<i>Some Essentials</i>	127
2.1	Unavoidable Biases Affecting IMF Studies	130
2.2	Discretizing an IMF: Optimal Sampling and the $m_{\max} - M_{\text{ecl}}$ Relation	131
2.3	Discretizing an IMF: Random Sampling and the Mass-Generating Function	135
2.4	A Practical Numerical Formulation of the IMF	136
2.5	Statistical Treatment of the Data	139
2.6	Binary Systems	141
3	<i>The Maximum Stellar Mass</i>	144
3.1	On the Existence of a Maximum Stellar Mass	144
3.2	The Upper Physical Stellar Mass Limit	145

3.3	The Maximal Stellar Mass in a Cluster, Optimal Sampling and Saturated Populations	147
3.3.1	Theory	147
3.3.2	Observational data	149
3.3.3	Interpretation	150
3.3.4	Stochastic or Regulated Star Formation?	151
3.3.5	A Historical Note	152
3.4	Caveats	153
4	<i>The Isolated Formation of Massive Stars</i>	153
5	<i>The IMF of Massive Stars</i>	157
6	<i>The IMF of Intermediate-Mass Stars</i>	159
7	<i>The IMF of Low-Mass Stars (LMSs)</i>	160
7.1	Galactic-Field Stars and the Stellar Luminosity Function	160
7.2	The Stellar Mass–Luminosity Relation	162
7.3	Unresolved Binary Stars and the Solar-Neighborhood IMF	166
7.4	Star Clusters	169
8	<i>The IMF of Very Low-Mass Stars (VLMSs) and of Brown Dwarfs (BDs)</i>	177
8.1	BD and VLMS Binaries	178
8.2	The Number of BDs per Star and BD Universality	181
8.3	BD Flavors	182
8.4	The Origin of BDs and Their IMF	184
9	<i>The Shape of the IMF from Resolved Stellar Populations</i>	186
9.1	The Canonical, Standard or Average IMF	187
9.2	The IMF of Systems and of Primaries	190
9.3	The Galactic-Field IMF	191
9.4	The Alpha Plot	191
9.5	The Distribution of Data Points in the Alpha-Plot	194
10	<i>Comparisons and Some Numbers</i>	197
10.1	The Solar-Neighborhood Mass Density and Some Other Numbers	197
10.2	Other IMF Forms and Cumulative Functions	198
11	<i>The Origin of the IMF</i>	199
11.1	Theoretical Notions	201
11.2	The IMF from the Cloud-Core Mass Function?	206
12	<i>Variation of the IMF</i>	209
12.1	Trivial IMF Variation Through the $m_{\max} - M_{\text{ecl}}$ Relation	209
12.2	Variation with Metallicity	210
12.3	Cosmological Evidence for IMF Variation	211
12.4	Top-Heavy IMF in Starbursting Gas	212

12.5	Top-Heavy IMF in the Galactic Center	213
12.6	Top-Heavy IMF in Some Star-Burst Clusters	214
12.7	Top-Heavy IMF in Some Globular Clusters (GCs)	215
12.8	Top-Heavy IMF in UCDS	218
12.9	The Current State of Affairs Concerning IMF Variation with Density and Metallicity and Concerning Theory	221
13	<i>Composite Stellar Populations: The IGIMF</i>	224
13.1	IGIMF Basics	225
13.2	IGIMF Applications, Predictions and Observational Verification	227
14	<i>The Universal Mass Function</i>	232
15	<i>Concluding Comments</i>	233
	<i>Acknowledgments</i>	235
	<i>References</i>	235

Abstract: The current knowledge on the stellar IMF is documented. It is usually described as being invariant, but evidence to the contrary has emerged: it appears to become top-heavy when the star-formation rate density surpasses about $0.1 M_{\odot}/(\text{year pc}^3)$ on a pc scale and it may become increasingly bottom-heavy with increasing metallicity and in increasingly massive elliptical galaxies. It declines quite steeply below about $0.07 M_{\odot}$ with brown dwarfs (BDs) and very low mass stars having their own IMF. The most massive star of mass m_{max} formed in an embedded cluster with stellar mass M_{ecl} correlates strongly with M_{ecl} being a result of gravitation-driven but resource-limited growth and fragmentation-induced starvation. There is no convincing evidence whatsoever that massive stars do form in isolation. Massive stars form above a density threshold in embedded clusters which become *saturated* when $m_{\text{max}} = m_{\text{max}^*} \approx 150 M_{\odot}$ which appears to be the canonical physical upper mass limit of stars. Super-canonical massive stars arise naturally due to stellar mergers induced by stellar-dynamical encounters in binary-rich very young dense clusters.

Various methods of discretising a stellar population are introduced: *optimal sampling* leads to a mass distribution that perfectly represents the exact form of the desired IMF and the $m_{\text{max}} - M_{\text{ecl}}$ relation, while *random sampling* results in statistical variations of the shape of the IMF. The observed $m_{\text{max}} - M_{\text{ecl}}$ correlation and the small spread of IMF power-law indices together suggest that optimally sampling the IMF may be the more realistic description of star formation than random sampling from a universal IMF with a constant upper mass limit.

Composite populations on galaxy scales, which are formed from many pc scale star formation events, need to be described by the integrated galactic IMF. This IGIMF varies systematically from top-light to top-heavy in dependence of galaxy type and star formation rate, with dramatic implications for theories of galaxy formation and evolution.

1 Introduction and Historical Overview

The distribution of stellar masses that form together, the stellar initial mass function (IMF), is one of the most important astrophysical distribution functions. The determination of the IMF is a very difficult problem because stellar masses cannot be measured directly and because observations usually cannot assess all stars in a population requiring elaborate bias corrections. Indeed, the stellar IMF is not measurable (the IMF Unmeasurability Theorem on p. 129). Nevertheless, impressive advances have been achieved such that the shape of the IMF is reasonably well understood from low-mass brown dwarfs (BDs) to very massive stars.

The IMF is of fundamental importance because it is a mathematical expression for describing the mass spectrum of stars born collectively in “one event.” Here, *one event* means a *gravitationally driven collective process of transformation of the interstellar gaseous matter into stars on a spatial scale of about 1 pc and within about 1 Myr*. Throughout this chapter, such events are referred to as *embedded star clusters*, but they need not lead to gravitationally bound long-lived open or globular clusters. Another astrophysical function of fundamental importance is the star-formation history (SFH) of a stellar system.

The IMF and the SFH are connected through complex self-regulating physical processes on galactic scales, whereby it can be summarized that for late-type galaxies, the star-formation rate (SFR) increases with increasing galaxy mass and the deeper gravitational potential. For early-type galaxies, the same is true except that the SFH was of short duration (\lesssim few Gyr). Together, the IMF and SFH contain the essential information on the transformation of dark gas to shining

stars and the spectral energy distribution thereof. They also contain the essential information on the cycle of matter, which fraction of it is locked up in feeble stars and substellar objects and how much of it is returned enriched with higher chemical elements to the interstellar medium or atmosphere of a galaxy. Knowing the rate with which matter is converted to stars in galaxies is essential for understanding the matter cycle and the matter content in the universe at a fundamental level.

This chapter is meant to outline essentials in IMF work, to document its form which appears to be invariant for the vast majority of resolved star-formation events, and to describe modern evidence for IMF variation and how whole galaxies are to be described as composite or complex populations. The literature on the IMF is vast, and it is unfortunately not possible to cover every research paper on this topic, although some attempt has been made to be as inclusive as possible. [Section 1](#) gives a brief historical review and a short overview of the topic, and pointers to other reviews are provided in [Sect. 1.8](#).

1.1 Solar Neighborhood

Given the importance of the IMF, a major research effort has been invested to distill its shape and variability. It began by first considering the best-known stellar sample, namely, that in the neighborhood of the Sun.

The seminal contribution by Salpeter (1955) while staying in Canberra first described the IMF as a power law, $dN = \xi(m) dm = k m^{-\alpha}$, where dN is the number of stars in the mass interval $m, m + dm$ and k is the normalization constant. By modeling the spatial distribution of the then observed stars with assumptions on the star-formation rate, Galactic-disk structure and stellar evolution timescales, Salpeter arrived at the power-law index (or “slope”) $\alpha = 2.35$ for $0.4 \lesssim m/M_{\odot} \lesssim 10$, which today is known as the “Salpeter IMF”¹

This IMF form implies a diverging mass density for $m \rightarrow 0$, which was interesting since dark matter was speculated, until the early 1990s, to possibly be made up of faint stars or substellar objects. Studies of the stellar velocities in the solar neighborhood also implied a large amount of missing, or dark, mass in the disk of the Milky Way (MW) (Bahcall 1984). Careful compilation in Heidelberg of the Gliese *Catalogue of Nearby Stars* beginning in the 1960s (Jahreiß and Wielen 1997)² and the application at the beginning of the 1980s of an innovative photographic pencil-beam survey technique reaching deep into the Galactic field in Edinburgh by Reid and Gilmore (1982) significantly improved knowledge of the space density of low-mass stars (LMSs, $m/M_{\odot} \lesssim 0.5$).

Major studies extending Salpeter’s work to lower and larger masses followed, showing that the mass function (MF) of Galactic-field stars turns over below one solar mass thus avoiding the divergence. Since stars with masses $m \lesssim 0.8 M_{\odot}$ do not evolve significantly over the age of the Galactic disk, the MF equals the IMF for these. While the work of Miller and Scalo (1979) relied on using the then-known nearby stellar sample to define the IMF for $m < 1 M_{\odot}$, Scalo (1986) relied mostly on a more recent deep pencil-beam star-count survey using photographic

¹As noted by Zinnecker (2011), Salpeter used an age of 6 Gyr for the MW disk; had he used the now adopted age of 12 Gyr, he would have arrived at a “Salpeter index” $\alpha \approx 2.05$ instead of 2.35.

²The latest version of the catalogue can be found at <http://www.ari.uni-heidelberg.de/datenbanken/aricns/>, while <http://www.nstars.nau.edu/> contains the Nearby Stars (NStars) data base.

plates. Scalo (1986) stands out as the most thorough and comprehensive analysis of the IMF in existence, laying down notation and ideas in use today.

The form of the IMF for low-mass stars was revised in the early 1990s in Cambridge, especially through the quantification of significant nonlinearities in the stellar mass–luminosity relation and evaluation of the bias due to unresolved binary systems (Kroupa et al. 1990, 1991, 1993). This work led to a detailed understanding of the shape of the stellar luminosity function (LF) in terms of stellar physics. It also resolved the difference between the results obtained by Miller and Scalo (1979) and Scalo (1986) through rigorous modeling of all biases affecting local trigonometric-based and distant photometric-parallax-based surveys, such as come from an intrinsic metallicity scatter, evolution along the main sequence, and contraction to the main sequence. In doing so, this work also included an updated local stellar sample *and* the then best-available deep pencil-beam survey. As such, it stands unique today as being the only rigorous analysis of the late-type-star MF using simultaneously both the *nearby trigonometric parallax* and the *far, pencil-beam* star-count data to constrain the one underlying MF of stars. This study was further extended to an analysis of other ground-based pencil-beam surveys being in excellent agreement with measurements with the HST of the LF through the entire thickness of the MW (● Fig. 4-9 below).

These results ($\alpha \approx 1.3$, $0.1\text{--}0.5 M_{\odot}$) were confirmed by Reid et al. (2002) using updated local star counts that included Hipparcos parallax data. Indeed, the continued, long-term observational effort on nearby stars by Neill Reid, John Gizis and collaborators forms one of the very major pillars of modern IMF work; continued discussion of controversial interpretations has much improved and sharpened our general understanding of the issues. A reanalysis of the nearby mass function of stars in terms of a log-normal form in the mass range $0.07\text{--}1 M_{\odot}$ was provided by Gilles Chabrier, finding agreement to the deep HST star-count data once unresolved multiple stars and a metal-deficient color-magnitude relation for thick-disk M dwarfs are accounted for (Chabrier 2003b). The log-normal form with a power-law extension to high masses is indistinguishable from the older canonical two-part power-law form (see ● Fig. 4-24 below). The immense analysis by Bochanski et al. (2010) of the LF and MF of 15×10^6 field low-mass dwarfs ($0.1 \lesssim m/M_{\odot} \lesssim 0.8$) derived from Sloan Digital Sky Survey Data Release 6 photometry using the photometric parallax method again finds good consistency with the previous work on the stellar LF and MF.

The Galactic-field stellar IMF for $0.08 \lesssim m/M_{\odot} \lesssim 1$ can thus be regarded as being reasonably well constrained. It converges to a finite mass density such that low-mass stars cannot be a dynamically significant contributor to the MW disk. The dynamical evidence for significant amounts of dark matter in the disk was finally eroded by Kuijken and Gilmore (1991) and Flynn and Fuchs (1994).

1.2 Star Clusters

In contrast to the Galactic-field sample, where stars of many ages and metallicities are mixed, clusters offer the advantage that the stars have the same age and metallicity and distance. And so a very large effort has been invested to try to extract the IMF from open and embedded clusters as well as from associations.

On the theoretical side, the ever-improving modeling of stellar and BD atmospheres being pushed forward with excellent results notably by the Lyon group (Isabelle Baraffe and Gilles Chabrier) has allowed consistently better constraints on the faint-star MF by a wide variety of

observational surveys.³ Furthermore, the development of high-precision N -body codes that rely on complex mathematical and algorithmic regularization of the equations of motions through the work of Sverre Aarseth and Seppo Mikkola (Aarseth 1999) and others has led to important progress on understanding the variation of the dynamical properties⁴ of stellar populations in individual clusters and the Galactic field.

In general, the MF found for clusters is consistent with the Galactic-field IMF for $m < 1 M_{\odot}$, but it is still unclear as to why open clusters have a significant deficit of white dwarfs (Fellhauer et al. 2003). Important issues are the rapid and violent early dynamical evolution of clusters due to the expulsion of residual gas and the associated loss of a large fraction of the cluster population and the density-dependent disruption of primordial binary systems (Kroupa et al. 2001).

Young clusters have thus undergone a highly complex dynamical evolution which continues into old age (Baumgardt and Makino 2003) and are therefore subject to biases that can only be studied effectively with full-scale N -body methods, thus imposing a complexity of analysis that surpasses that for the Galactic-field sample. The pronounced deficit of BDs and low-mass stars in the 600-Myr-old Hyades are excellent observational proof on how dynamical evolution affects the MF (Bouvier et al. 2008; Röser et al. 2011). While estimating masses of stars younger than a few Myr is subject to major uncertainty,³ differential reddening is also a complicated problem to handle when studying the IMF (► Sect. 2.1).

Important Note

In order to constrain the shape and variation of the IMF from star-count data in clusters, high-precision N -body modeling must be mastered in addition to stellar-evolution theory and knowledge of the properties of multiple stars.

1.3 Intermediate-Mass and Massive Stars

Intermediate-mass and particularly massive stars are rare, and so larger spatial volumes need to be surveyed to assess their distribution by mass.

For intermediate-mass and massive stars, the Scalo (1986) IMF is based on a combination of Galactic-field star counts and OB association data and has a slope $\alpha \approx 2.7$ for $m \gtrsim 2 M_{\odot}$ with much uncertainty for $m \gtrsim 10 M_{\odot}$. The previous determination by Miller and Scalo (1979) also implied a relatively steep field-IMF with $\alpha \approx 2.5$, $1 \leq m/M_{\odot} \leq 10$ and $\alpha \approx 3.3$, $m > 10 M_{\odot}$. Elmegreen and Scalo (2006) point out that structure in the MF is generated at a mass if the SFR of the population under study varies on a timescale comparable to the stellar evolution

³Here, it should be emphasized and acknowledged that the intensive and highly fruitful discourse between Guenther Wuchterl and the Lyon group has led to the important understanding that the classical evolution tracks computed in Lyon and by others are unreliable for ages less than a few Myr (Tout et al. 1999; Wuchterl and Tscharnuter 2003). This comes about because the emerging star's structure retains a memory of its accretion history. In particular, Wuchterl and Klessen (2001) present a SPH computation of the gravitational collapse and early evolution of a solar-type star documenting the significant difference to a pre-main sequence track if the star is instead classically assumed to form in isolation.

⁴The *dynamical properties* of a stellar system are its *mass* and, if it is a multiple star, its orbital parameters (*semimajor axis*, *mass ratio*, *eccentricity*, both inner and outer if it is a higher-order multiple system). See also ► Sect. 2.6.

timescale of its stars near that mass. Artificial structure in the stellar IMF may be deduced in this case, and this effect is particularly relevant for the field IMF. Perhaps the peculiar structure detected by Gouliermis et al. (2005) in the field MF of stars in the Large Magellanic Cloud may be due to this effect.

An interesting finding in this context reported by Massey (1998, 2003) is that the OB stellar population found in the field of the LMC has a very steep MF, $\alpha \approx 4.5$, which can be interpreted to be due to the preferred formation of small groups or even isolated O- and B-type stars (e.g., Selier et al. 2011). Another or an additional effect influencing the deduced shape of the IMF via a changing SFR (Elmegreen and Scalo 2006) is the dynamical ejection of OB stars from dynamically unstable cores of young clusters. This may lead to such a steep IMF because, as dynamical work suggests, preferentially, the less-massive members of a core of massive stars are ejected (Clarke and Pringle 1992; Pflamm-Altenburg and Kroupa 2006). This process would require further study using fully consistent and high-precision N -body modeling of young clusters to see if the observed distribution of field-OB stars can be accounted for with this process alone or if indeed an exotic star-formation mode needs to be invoked to explain some of the observations.

The IMF for intermediate-mass and massive stars deduced from star counts in the field of a galaxy may therefore yield false information on the true shape of the stellar IMF, and stellar samples with well-defined formation histories such as OB associations and star clusters are more useful to constrain the IMF. Photometric surveys of such regions, while essential for reaching faint stars, do not allow a reliable assessment of the mass distribution of massive stars since their spectral energy distribution is largely output at short wavelengths beyond the optical. Spectroscopic classification is therefore an essential tool for this purpose.

Phil Massey's work at Tucson, based on extensive spectroscopic classification of stars in OB associations and in star clusters, demonstrated that for massive stars $\alpha = 2.35 \pm 0.1$, $m \gtrsim 10 M_{\odot}$ (Massey 1998), for a large variety of physical environments as found in the MW and the Magellanic Clouds, namely, OB associations and dense clusters and for populations with metallicity ranging from near-solar abundance to about one tenth metal abundance (► Fig. 4-8 below). The significant differences in the metallicity between the outer and the inner edge of the Galactic disk seem not to influence star-formation, as Yasui et al. (2008) find no apparent difference in the stellar MFs of clusters in the extreme outer Galaxy compared to the rest of the disk down to $0.1 M_{\odot}$.

1.4 The Invariant IMF and Its Conflict with Theory

The thus empirically constrained stellar IMF can be described well by a two-part power-law form with $\alpha_1 = 1.3$ for $m \lesssim 0.5 M_{\odot}$ and with the *Salpeter/Massey index*, $\alpha_2 = 2.35$ for $m \gtrsim 0.5 M_{\odot}$, being remarkably invariant.⁵ This IMF is referred to as the *canonical IMF* and is given in ► Sect. 9.1 as the simpler two-part power-law form by (► 4.55) and as the log-normal plus power-law form by (► 4.56). The empirical result leads to the statement of the following hypothesis:

⁵Note that Scalo (1998) emphasizes that the IMF remains poorly constrained owing to the small number of massive stars in any one sample. This is a true albeit conservative standpoint, and the present authors prefer to accept Massey's result as a working IMF Universality Hypothesis.

Invariant IMF Hypothesis

There exists a universal parent distribution function which describes the distribution of stellar masses in individual star-forming events.

The *Invariant IMF Hypothesis* needs to be tested with star-count data in galactic fields and in individual clusters and OB associations for possible significant deviations. But it is mandatory to take into account the biases listed in [Sect. 2.1](#) when doing so.

The *Invariant IMF Hypothesis* is not consistent with theory (see also [Sects. 11.1](#) and [12](#)). There are two broad theoretical ansatzes for the origin of stellar masses:

The Jeans-Mass Ansatz: According to the Jeans-mass argument (e.g., Jeans 1902; Larson 1998; Bate and Bonnell 2005; Bonnell et al. 2006, ([4.50](#)) below), star formation at lower metallicity ought to produce stars of, on average, heavier mass and thus an effectively top-heavy IMF (i.e., the ratio between the number of massive stars and low mass stars ought to increase). At lower metallicity, the cooling is less efficient causing larger Jeans masses as a requirement for gravitational collapse to a protostar and thus larger stellar masses. That warmer gas produces an IMF shifted to larger masses has been demonstrated with state-of-the art SPH simulations (e.g., Klessen et al. 2007).

The Self-Regulatory Ansatz: Another approach is formulated by Adams and Fatuzzo (1996) who argue that the Jeans-Mass Ansatz is invalid since there is no preferred Jeans mass in a turbulent molecular cloud. Instead, they invoke the central-limit theorem⁶ together with self-regulated assembly, and they suggest that the final stellar masses are given by the balance between feedback energy from the forming star (accretion luminosity, outflows) and the rate of accretion from the proto-stellar envelope and circum-stellar disk. As the protostar builds up its luminosity increases until the accretion is shut off. When shutoff occurs depends on the accretion rate. Indeed, Basu and Jones (2004) explain the observed power-law extension of the IMF at large stellar masses as being due to a distribution of different accretion rates. The self-regulating character of star formation has been studied profusely by the group around Christopher McKee (e.g., Tan et al. 2006) and has been shown to lead to decreasing star-formation efficiencies with increasing metallicity (Dib et al. 2010, 2011; Dib 2011). In low-metallicity gas, the coupling of the photons to the gas is less efficient, causing a less effective opposition against the accreting material. And, at lower metallicity, the cooling is reduced, causing a higher temperature of the gas and thus a higher speed of sound with a larger accretion rate. The final stellar IMF is expected to be populated by more massive stars in metal-poor environments.

Both approaches can be refined by studying a distribution of physical conditions in a given star-forming cloud, but both lead to the same conclusion, namely, that low metallicity and

⁶Citing from Basu and Jones (2004), “According to the central limit theorem of statistics, if the mass of a protostellar condensation $M_c = f_1 \times f_2 \times \dots \times f_N$, then the distribution of M_c tends to a lognormal regardless of the distributions of the individual physical parameters $f_i (i = 1, \dots, N)$, if N is large. Depending on the specific distributions of the f_i , a convergence to a lognormal may even occur for moderate N .” The central limit theorem was invoked for the first time by Zinnecker (1984) to study the form of the IMF from hierarchical fragmentation of collapsing cloud cores.

high temperature ought to produce top-heavy stellar IMFs. This leads to the following robust theoretical IMF result:

The Variable IMF Prediction

Both the Jeans-mass and the self-regulation arguments invoke very different physical principles, and yet they lead to the same result: The IMF ought to become top-heavy under low-metallicity and high-temperature star-forming conditions.

Star formation in the very early universe must have therefore produced top-heavy IMFs (Bromm et al. 2001; Clark et al. 2011). But the samples of simple-stellar populations spanning all cosmological epochs (globular clusters to current star-formation in embedded clusters) available in the Local Group of galaxies have until recently not shown convincing evidence supporting The Variable-IMF Prediction. This issue is addressed in more detail in [Sect. 12.9](#).

1.5 Philosophical Note

Much of the current discussion on star formation, from the smallest to the largest (galaxy-wide) scales, can be categorized into two broad conceptual approaches which are related to the Jeans-Mass Ansatz versus the Self-Regulatory Ansatz of [Sect. 1.4](#):

Approach A is related to the notion that star formation is inherently stochastic such that the IMF is a probabilistic distribution function only. This is a natural notion under the argument that the processes governing star-formation are so many and complex that the outcome is essentially stochastic in nature. Followers of this line of reasoning argue for example that massive stars can form in isolation and that the mass of the most massive star cluster forming in a galaxy depends on the time-scale over which an ensemble of star clusters is considered (the *size-of-sample effect*, even at very low SFRs a galaxy would produce a very massive star cluster if one waits long enough, i.e., if the sample of clusters is large enough). Approach A can be formulated concisely as *nature plays dice when stars form*.

Approach B is related to the notion that nature is inherently self-regulated and deterministic. This is a natural notion given that physical processes must always depend on the boundary conditions which are a result of the physical processes at hand. An example of such would be gravitationally-driven growth processes with feedback in media with limited resources. The emerging phenomena such as the distribution of stellar masses, of star-cluster masses and/or of how phase-space is populated to make binary stellar systems are, at least in principle, computable. They are computable in the sense that statistical mathematics provides the required tools such that the distribution functions used to describe the outcomes are subject to constraints. For example, a young stellar population of mass M_{ecl} is excellently described by $\xi(m)$ with the condition $m \leq m_{\text{max}}(M_{\text{ecl}})$. However, purely random sampling from $\xi(m)$ even under this constraint will not reproduce a realistic population if nature follows Optimal Sampling (p. 132). This is because Optimal Sampling will never allow a cluster to be made up of $M_{\text{ecl}}/m_{\text{max}}$ stars of mass m_{max} , while constrained random sampling would. Approach B can be formulated concisely as *nature does not play dice when stars form*.

Depending on which of the two notions is applied, the resulting astrophysical description of galaxies leads to very diverging results. Either a galaxy can be described as an object in which stars form purely stochastically such that the galaxy-wide IMF is equal to the stellar IMF (Approach A). In this case, a thousand small groups of 20 pre-main sequence stars will have the same stellar IMF as one very young star cluster containing 20,000 stars, or an embedded cluster or a galaxy is understood to be a highly self-regulated system such that the galaxy-wide IMF differs from the stellar IMF (Approach B, [Sect. 13.1](#)). According to this notion, a thousand small groups of 20 pre-main sequence stars would not contain a single star with $m > 5 M_{\odot}$, while a very young star cluster of 20,000 stars would contain many such stars. The different approaches have very different implications for understanding the matter cycle in the universe.

1.6 Hypothesis Testing

The studies aimed at constraining the stellar IMF observationally typically have the goal of testing the Invariant IMF Hypothesis (p. 123) either in individual star-forming events such as in a star clusters or on galaxy-wide scales. Here, it is important to be reminded of the following:

Elementary Logics of Hypothesis Testing

Negation of a hypothesis I does not imply that the alternative hypothesis II is correct.

By showing that hypothesis I is consistent with some data does *not* imply that an alternative hypothesis II is therewith ruled out. A case in point is the discussion about dark matter and dark energy: Ruling out the standard cosmological (LCDM) model does not imply that any particular alternative is correct (Kroupa et al. 2010; Kroupa 2012). Conversely, ruling out a particular alternative does not imply that the LCDM model is correct (Wojtak et al. 2011).

Concerning the IMF, if a purely stochastic model (approach A, [Sect. 1.5](#)) is consistent with some observational data, then this does not imply that the alternative (optimal sampling, which is related to approach B) is falsified.

A case in point is provided by the following example relevant for the tests of the IGIMF theory on p. 231: The masses of an ensemble of observed dwarf galaxies are calculated from spectral energy distribution modeling using the universal canonical IMF. These masses are then applied in testing a possible variation of the galaxy-wide IMF in terms of the UV and H α flux ratios. If the universal IMF calculations allow for fluctuating SFRs whereas the variable IMF calculations do not, then the (wrong) conclusion of such an approach would plausibly be that nature appears to play dice because observational data naturally contain measurement uncertainties which act as randomization agents.

The consistent approach would instead be to compute the galaxy masses assuming a variable galaxy-wide IMF to test whether the hypothesis that the IMF varies systematically with galaxy mass can be discarded. The logically consistent procedure would be to calculate all fluxes within both scenarios independently of each other and assuming in both that the SFR can fluctuate and to test these calculations against the observed fluxes. The result of this consistent procedure is opposite to those of the above inconsistent procedure in that the data are in better agreement with the systematically variable IMF, that is, that nature does not play dice.

A final point to consider is when a hypothesis ought to be finally discarded. Two examples illustrate this: Consider the Taurus-Auriga and Orion star-forming clouds. Here, the number of stars with $m \gtrsim 1 M_{\odot}$ is significantly below the expectation from the purely stochastic model (see box IGIMF predictions/tests on p. 231). This unambiguously falsifies the stochastic model. But the data are in excellent agreement with the expectation from the IGIMF theory. Is it then meaningful to nevertheless keep adopting the stochastic model on cluster and galaxy problems? Another example, being perhaps more relevant to [Sect. 1.5](#), is the issue with the current standard cosmological model. It is ruled out by Peebles and Nusser (2010) and Kroupa (2012). Should it nevertheless be adopted in further cosmological and related research?

1.7 About This Text

As is evident from the above introduction, the IMF may be well defined and is quite universal in each star-forming event out of which comes a spatially and temporally well-correlated set of stars. But many such events will produce a summed IMF which may be different because the individual IMFs need to be added whereby the distribution of the star-formation events in mass, space, and time becomes an issue. It thus emerges that it is necessary to distinguish between simple stellar populations and composite populations. Some definitions are useful:

Definitions

- A *simple population* obtains from a spatially (\lesssim few pc) and temporarily (\lesssim Myr) correlated star-formation event (CSFE, also referred to as a *collective star formation event*, being essentially an embedded star cluster). The mass of a CSFE may range from a few solar masses (a few binary stars) upwards.
- A *composite or complex population* consists of more than one simple population.
- The *stellar IMF* refers to the IMF of stars in a simple population.
- A *composite or integrated IMF* is the IMF of a composite or complex population, that is, a population composed of many CSFEs, most of which may be gravitationally unbound. The galaxy-wide version is the IGIMF ([Sect. 13.1](#)).
- The PDMF is the present-day MF of a stellar population not corrected for stellar evolution nor for losses through stellar deaths. Note that a *canonical PDMF* is a PDMF derived from a canonical IMF.
- A *stellar system* can be a multiple star or a single star. It has *dynamical properties* (footnote 4 on p. 121).
- The *system luminosity* or *mass function* is the LF or MF obtained by counting all stellar systems. The *stellar LF* or *stellar MF* is the true distribution of all stars in the sample, thereby counting all individual companions in multiple systems. This is also referred to as the *individual-star LF/MF*.
- Note $1 \text{ km/s} = 1.0227 \text{ pc/Myr}$, $1 \text{ g cm}^{-3} = 1.478 \times 10^{22} M_{\odot} \text{ pc}^{-3}$, and $1 \text{ g cm}^{-2} = 4788.4 M_{\odot} \text{ pc}^{-2}$ for a solar-metallicity gas.
- The following additional abbreviations are used: SFR = star formation rate in units of $M_{\odot} \text{ year}^{-1}$; SFH = star formation history = SFR as a function of time; SFRD = star formation rate density in units of $M_{\odot} \text{ year}^{-1} \text{ pc}^{-3}$.

This treatise provides an overview of the general methods used to derive the IMF with special attention on the pitfalls that are typically encountered. The binary properties of stars and of brown dwarfs are discussed as well because they are essential to understand the true shape of the stellar IMF. While the stellar IMF appears to have emerged as being universal in star-formation events as currently found in the Local Group of galaxies, the recent realization that star clusters limit the mass spectrum of their stars is one form of IMF variation and has interesting implications for the formation of stars in a cluster and leads to the insight that composite populations must show IMFs that differ from the stellar IMF in each cluster. With this finale, this treatise reaches the cosmological arena.

1.8 Other IMF Reviews

The seminal contribution by Scalo (1986) on the IMF remains a necessary source for consultation on the fundamentals of the IMF problem. The landmark review by Massey (2003) on massive stars in the Local Group of galaxies is an essential read, as is the review by Zinnecker and Yorke (2007) on massive-star formation. Other reviews of the IMF are by Scalo (1998), Kroupa (2002), Chabrier (2003a), Bonnell et al. (2007), Elmegreen (2009), and Bastian et al. (2010). The proceedings of the “38th Herstmonceux Conference on the Stellar Initial Mass Function” (Gilmore and Howell 1998) and the proceedings of the “IMF50” meeting in celebration of Ed Salpeter’s 80th birthday (Corbelli et al. 2005) contain a wealth of important contributions to the field. A recent major but also somewhat exclusive conference on the IMF was held from June 20–25, 2010, in Sedona, Arizona, for researchers to discuss the recently accumulating evidence for IMF variations: “UP2010: Have Observations Revealed a Variable Upper End of the Initial Mass Function?” The published contributions are a unique source of information on this problem (Treyer et al. 2011). A comprehensive review of extreme star formation is available by Turner (2009).

2 Some Essentials

Assuming the relevant biases listed in [Sect. 2.1](#) have been corrected for such that all binary and higher-order stellar systems can be resolved into individual stars in some complete population such as the solar neighborhood and that only main-sequence stars are selected for, then the number of single stars per pc^3 in the mass interval m to $m + dm$ is $dN = \Xi(m) dm$, where $\Xi(m)$ is the *present-day mass function* (PDMF). The number of single stars per pc^3 in the absolute P-band magnitude interval M_P to $M_P + dM_P$ is $dN = -\Psi(M_P) dM_P$, where $\Psi(M_P)$ is the stellar luminosity function (LF) which is constructed by counting the number of stars in the survey volume per magnitude interval and P signifies an observational photometric passband such as the V- or I-band. Thus,

$$\Xi(m) = -\Psi(M_P) \left(\frac{dm}{dM_P} \right)^{-1}. \quad (4.1)$$

Note that the minus sign comes in because increasing mass leads to decreasing magnitudes and that the LF constructed in one photometric passband P can be transformed into another band P' by

$$\Psi(M_P) = \frac{dN}{dM_{P'}} \frac{dM_{P'}}{dM_P} = \Psi(M_{P'}) \frac{dM_{P'}}{dM_P} \quad (4.2)$$

if the function $M_{p'} = \text{fn}(M_p)$ is known. Such functions are equivalent to color–magnitude relations.

Since the derivative of the stellar mass–luminosity relation (MLR), $m(M_p) = m(M_p, Z, \tau, \mathbf{s})$, enters the calculation of the MF, any uncertainties in stellar structure and evolution theory on the one hand (if a theoretical MLR is relied upon) or in observational ML-data on the other hand will be magnified accordingly. This problem cannot be avoided if the mass function is constructed by converting the observed stellar luminosities one by one to stellar masses using the MLR and then binning the masses because the derivative of the MLR nevertheless creeps in through the binning process because *equal luminosity intervals are not mapped into equal mass intervals*. The dependence of the MLR on the star’s chemical composition, Z , its age, τ , and spin vector \mathbf{s} is explicitly stated here since stars with fewer metals than the Sun are brighter (lower opacity), main-sequence stars brighten with time and loose mass, and rotating stars are dimmer because of the reduced internal pressure. Mass loss and rotation are significant factors for intermediate and especially high-mass stars (Penny et al. 2001).

The IMF, or synonymously here the IGIMF (☛ Sect. 13.1), follows by correcting the observed number of main-sequence stars for the number of stars that have evolved off the main sequence. Defining $t = 0$ to be the time when the Galaxy that now has an age $t = \tau_G$ began forming, the number of stars per pc^3 in the mass interval $m, m + dm$ that form in the time interval $t, t + dt$ is $dN = \xi(m; t) dm b'(t) dt$, where the expected time dependence of the IMF is explicitly stated (☛ Sect. 13.1) and where $b'(t) = b(t)/\tau_G$ is the normalized star-formation history, $(1/\tau_G) \int_0^{\tau_G} b(t) dt = 1$. Stars that have main-sequence lifetimes $\tau(m) < \tau_G$ leave the stellar population unless they were born during the most recent time interval $[\tau_G - \tau(m), \tau_G]$. The number density of such stars still on the main sequence with initial masses computed from their present-day masses and their ages in the range $m, m + dm$ and the total number density of stars with $\tau(m) \geq \tau_G$ are, respectively,

$$\Xi(m) = \xi(m) \frac{1}{\tau_G} \times \begin{cases} \int_{\tau_G - \tau(m)}^{\tau_G} b(t) dt & , \quad \tau(m) < \tau_G, \\ \int_0^{\tau_G} b(t) dt & , \quad \tau(m) \geq \tau_G, \end{cases} \quad (4.3)$$

where the time-averaged IMF, $\xi(m)$, has now been defined. Thus, for low-mass stars $\Xi = \xi$, while for a subpopulation of massive stars that has an age $\Delta t \ll \tau_G$, $\Xi = (\Delta t/\tau_G) \xi$ for those stars of mass m for which the main-sequence lifetime $\tau(m) > \Delta t$, indicating how an observed high-mass IMF in an OB association, for example, has to be scaled to the Galactic-field IMF for low-mass stars, assuming continuity of the IMF. In this case, the different spatial distribution via different disk-scale heights of old and young stars also needs to be taken into account, which is done globally by calculating the stellar surface density in the MW disk (Miller and Scalo 1979; Scalo 1986). Thus, we can see that joining the cumulative low-mass star counts to the snapshot view of the massive-star IMF is nontrivial and affects the shape of the IMF in the notorious mass range $\approx 0.8\text{--}3 M_\odot$, where the main-sequence lifetimes are comparable to the age of the MW disk (☛ Fig. 4-26, bottom panel). For a population in a star cluster or association with an age $\tau_{\text{cl}} \ll \tau_G$, τ_{cl} replaces τ_G in (☛ 4.3). Examples of the time modulation of the IMF are $b(t) = 1$ (constant star-formation rate) or a Dirac-delta function, $b(t) = \delta(t - t_0)$ (all stars formed at the same time t_0).

The stellar IMF can conveniently be written as an arbitrary number of power-law segments,

$$\xi_{\text{BD}} = k k_{\text{BD}} \left(\frac{m}{m_1} \right)^{-\alpha_0}, \quad (4.4)$$

$$\xi_{\text{star}}(m) = k \begin{cases} \left(\frac{m}{m_1}\right)^{-\alpha_1} & , \quad m_1 < m \leq m_2 & , \quad n = 1, \\ \left[\prod_{i=2}^{n \geq 2} \left(\frac{m_i}{m_{i-1}}\right)^{-\alpha_{i-1}}\right] \left(\frac{m}{m_n}\right)^{-\alpha_n} & , \quad m_n < m \leq m_{n+1} & , \quad n \geq 2, \end{cases} \quad (4.5)$$

where k_{BD} and k contain the desired scaling, $0.01 M_{\odot}$ is about the minimum mass of a BD (see footnote 15 on p. 187), and the mass ratios ensure continuity. Here, the separation into the IMF of BDs and of stars has already been explicitly stated (see [Sect. 8](#)).

Often used is the “logarithmic mass function” ([Table 4-3](#) below),

$$\xi_{\text{L}}(m) = (m \ln 10) \xi(m), \quad (4.6)$$

where $dN = \xi_{\text{L}}(m) dl m$ is the number of stars with mass in the interval $lm, lm + dl m$ ($lm \equiv \log_{10} m$).⁷

The stellar mass of an embedded cluster, M_{ecl} , can be used to investigate the expected number of stars above a certain mass m ,

$$N(> m) = \int_m^{m_{\text{max}^*}} \xi(m') dm', \quad (4.7)$$

with the mass in stars of the whole (originally embedded) cluster, M_{ecl} , being calculated from

$$M_{1,2} = \int_{m_1}^{m_2} m' \xi(m') dm', \quad (4.8)$$

with $M_{\text{ecl}} = M_{1,2}$ for $m_1 = m_{\text{low}} \approx 0.07 M_{\odot}$ (about the hydrogen burning mass limit) and $m_2 = m_{\text{max}^*} = \infty$ (the *Massey assertion*, p. 145, but see [Sect. 3.3](#)). There are two unknowns ($N(> m)$ and k) that can be solved for by using the two equations above.

It should be noted that the IMF is not a measurable quantity: Given that we are never likely to learn the exact dynamical history of a particular cluster or population, it follows that we can *never* ascertain the IMF for any individual cluster or population. This can be summarized concisely with the following theorem:

The IMF Unmeasurability Theorem

The IMF cannot be extracted directly for any individual stellar population.

Proof: For clusters younger than about 1 Myr, star formation has not ceased, and the IMF is therefore not assembled yet, and the cluster cores consisting of massive stars have already dynamically ejected members (Pflamm-Altenburg and Kroupa 2006). Massive stars ($m \gtrsim 30 M_{\odot}$) leave the main sequence before they are fully assembled (Maeder and Behrend 2002). For clusters with an age between 0.5 and a few Myr, the expulsion of residual gas has led to a loss of stars (Kroupa et al. 2001). Older clusters are either still losing stars due to residual gas expulsion or are evolving secularly through evaporation driven by energy equipartition (Vesperini and HEGGIE 1997; Baumgardt and Makino 2003). There exists thus no time when all stars are assembled in an observationally accessible volume (i.e., a star cluster). An observer is never able to access all phase-space variables of all potential members of an OB association. The field population is a mixture of many star-formation events whereby it can practically not be proven that a complete population has been documented. \square

⁷Note that Scalo (1986) calls $\xi_{\text{L}}(m)$ the *mass function* and $\xi(m)$ the *mass spectrum*.

Note that the IMF Unmeasurability Theorem implies that individual clusters cannot be used to make deductions on the similarity or not of their IMFs, unless a complete dynamical history of each cluster is available.

Notwithstanding this pessimistic theorem, it is nevertheless necessary to observe and study star clusters of any age. Combined with thorough and realistic N -body modeling, the data do lead to essential *statistical* constraints on the IMF Universality Hypothesis (p. 189, see also p. 123).

2.1 Unavoidable Biases Affecting IMF Studies

Past research has uncovered a long list of biases that affect the conversion of the observed distribution of stellar brightnesses to the underlying stellar IMF. These are just as valid today, and in particular, analysis of the GAIA-space-mission data will need to take the relevant ones into account before the stellar IMF can be constrained anew. The list of all unavoidable biases affecting stellar IMF studies is provided here with key references addressing these:

Malmquist bias (affects MW-field star counts): Stars of the same mass but with different ages, metallicities, and spin vectors have different colors and luminosities which lead to errors in distance measurements in flux-limited field star counts using photometric parallax (Stobie et al. 1989).

Color-magnitude relation (affects MW-field star counts): Distance measurements through photometric parallax are systematically affected if the true color-magnitude relation of stars deviates from the assumed relation (Reid and Gizis 1997) (but see footnote 14 on p. 162).

Lutz-Kelker bias (affects MW-field star counts): A distance-limited survey is affected by parallax measurement uncertainties such that the spatial stellar densities are estimated wrongly (Lutz and Kelker 1973). Correcting for this bias will be required when analyzing GAIA-space mission star-count data.

Unresolved multiple stars (affects all star counts): Companions to stars can be missed because their separation is below the resolution limit or because the companion's luminosity is below the flux limit (Kroupa et al. 1991). When using photometric parallax to determine distances and space densities, unresolved multiple systems appear nearer and redder. This affects the measured disk scale height as a function of stellar spectral type (Kroupa et al. 1993). Missed companions have a significant effect on the deduced shape of the IMF for $m \lesssim 1 M_{\odot}$ (Kroupa et al. 1991, 1993; Malkov and Zinnecker 2001) but do not significantly affect the shape of the stellar IMF for more massive stars (Maíz Apellániz and Úbeda 2005; Weidner et al. 2009).

Stellar mass-luminosity relation (MLR, affects all star counts): Main sequence stars of precisely the same chemical composition, age, and spins follow one perfect mass-luminosity relation. Its nonlinearities map a featureless stellar IMF to a structured LF, but theoretical MLRs are not reliable (Kroupa et al. 1990). An ensemble of field stars do not follow one stellar mass-luminosity relation such that the nonlinearities in it that map to structure in the stellar LF are smeared out (Kroupa et al. 1993). Correcting for this bias will be required when analyzing GAIA-space mission star-count data. Pre-main sequence stars have a complicated and time-varying mass-luminosity relation (Piskunov et al. 2004).

Varying SFH (affects all star counts): Variations of the SFH of a population under study over a characteristic time scale leads to structure in the deduced IMF at a mass scale at

which stars evolve on that time scale, if the observer wrongly assumes a constant SFH (Elmegreen and Scalo 2006).

Stellar evolution (affects all star counts): Present-day stellar luminosities must be transformed to initial stellar masses. This relies on stellar-evolution theory (Scalo 1986).

Binary-stellar evolution (affects all star counts): Present-day stellar luminosities must be transformed to initial stellar masses, but this may be wrong if the star is derived from an interacting binary. If important, then this only affects the massive-star IMF (F. Schneider and R. Izzard, private communication).

Pre-main sequence evolution (affects all populations with stars younger than a few 10^8 year): In young star clusters, the late-type stellar luminosities need to be corrected for the stars not yet being on the main sequence (e.g., Hillenbrand and Carpenter 2000). Pre-main sequence evolution tracks are highly uncertain for ages $\lesssim 1$ Myr (Tout et al. 1999; Wuchterl and Tscharnuter 2003). Field star-count data contain an admixture of young stars which bias the star counts (Kroupa et al. 1993).

Differential reddening (affects embedded star clusters): Patchily distributed gas and dust affects mass estimation. Variable extinction necessitates the introduction of an extinction limit which increases the lower mass limit to which the survey is complete (Andersen et al. 2009).

Binning: Deriving the IMF power-law index from a binned set of data is prone to significant bias caused by the correlation between the assigned weights and the number of stars per bin. Two solutions have been proposed: variable-sized binning (Maíz Apellániz and Úbeda 2005) and newly developed (effectively) bias-free estimators for the exponent and the upper stellar mass limit (Maschberger and Kroupa 2009).

Crowding (affects star counts in star clusters): A compact faraway star cluster can lead to crowding and superpositions of stars which affects the determination of the IMF systematically (Maíz Apellániz 2008).

Early and late star cluster evolution (affects star-counts in star clusters): A large fraction of massive stars are ejected from the cluster core skewing the MF in the cluster downward at the high-mass end (Pflamm-Altenburg and Kroupa 2006; Banerjee et al. 2012). When the residual gas is blown out of initially mass-segregated young clusters, they preferentially lose low-mass stars within a few Myr (Marks et al. 2008). Old star clusters evolve through energy equipartition driven evaporation of low-mass stars (Vesperini and Heggie 1997; Baumgardt and Makino 2003).

2.2 Discretizing an IMF: Optimal Sampling and the $m_{\max} - M_{\text{ecl}}$ Relation

In view of [Sects. 1.4](#) and [1.5](#), it is clearly necessary to be able to set up and to test various hypothesis as to how a stellar population emanates from a star-formation event. Two extreme hypotheses are (1) the stars born together are always perfectly distributed according to the stellar IMF and (2) the stars born together represent a random draw of masses from the IMF. Here, one method of perfectly distributing the stellar masses according to the form of the IMF is discussed. [Section 2.3](#) describes how to generate a random population of stars highly efficiently.

Note that throughout this chapter, the relevant physical quantity of a population is taken to be its mass and *never* the number of stars, N , which is not a physical quantity.

It is useful to consider the concept of optimally sampling a distribution function. The problem to be addressed is that there is a mass reservoir, M_{ecl} , which is to be distributed according to the IMF such that no gaps arise.

Ansatz: Optimal Sampling

Given a predefined form of a continuous distribution function, $\xi(m)$, of the variable $m \in [m_L, m_U]$ (Note: $m_U = m_{\text{max}^*}$ is adopted here for brevity of notation) such that $m_2 > m_1 \implies \xi(m_1) > \xi(m_2) > 0$, then the physical reservoir M_{ecl} is *optimally distributed* over $\xi(m)$ if the maximum available range accessible to m is covered with the condition that a m occurs once above a certain limit $m_{\text{max}} \in [m_L, m_U]$, $\int_{m_{\text{max}}}^{m_U} \xi(m) dm = 1$.

We define $\xi(m) = k p(m)$, where $p(m)$ is the density distribution function of stellar masses. The last statement in the above ansatz implies $k = 1 / (\int_{m_{\text{max}}}^{m_U} p(m) dm)$. Since the total mass in stars, $M_* = k \int_{m_L}^{m_{\text{max}}} m p(m) dm$, one obtains

$$M_* = \frac{\int_{m_L}^{m_{\text{max}}} m p(m) dm}{\int_{m_{\text{max}}}^{m_U} p(m) dm}.$$

It thus follows immediately that $m'_{\text{max}} > m_{\text{max}} \implies M_* > M_{\text{ecl}}$ and also $m'_{\text{max}} < m_{\text{max}} \implies M_* < M_{\text{ecl}}$. Thus, only $m'_{\text{max}} = m_{\text{max}} \implies M_* = M_{\text{ecl}}$. The concept of optimal sampling appears to be naturally related to how M_{ecl} is divided up among the stars: The largest chunk goes to m_{max} , and the rest is divided up hierarchically among the lesser stars (see Open Question II on p. 150).

The above ansatz can be extended to a discretized optimal distribution of stellar masses: Given the mass, M_{ecl} , of the population, the following sequence of individual stellar masses yields a distribution function which exactly follows $\xi(m)$,

$$m_{i+1} = \int_{m_{i+1}}^{m_i} m \xi(m) dm, \quad m_L \leq m_{i+1} < m_i, \quad m_1 \equiv m_{\text{max}}. \quad (4.9)$$

The normalization and the most massive star in the sequence are set by the following two equations:

$$1 = \int_{m_{\text{max}}}^{m_{\text{max}^*}} \xi(m) dm, \quad (4.10)$$

with

$$M_{\text{ecl}}(m_{\text{max}}) - m_{\text{max}} = \int_{m_L}^{m_{\text{max}}} m \xi(m) dm \quad (4.11)$$

as the closing condition. These two equations need to be solved iteratively. An excellent approximation is given by the following formula (equation 10 in Pflamm-Altenburg et al. 2007, assuming $m_{\text{max}^*} = 150 M_\odot$):

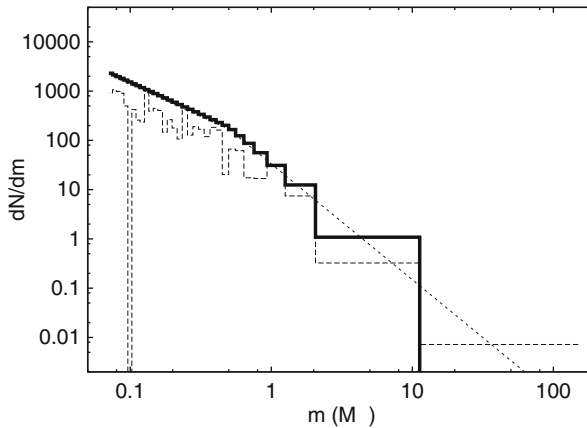
$$\log_{10} \left(\frac{m_{\text{max}}}{M_\odot} \right) = 2.56 \log_{10} \left(\frac{M_{\text{ecl}}}{M_\odot} \right) \left(3.82^{9.17} + \left[\log_{10} \left(\frac{M_{\text{ecl}}}{M_\odot} \right) \right]^{9.17} \right)^{-\frac{1}{9.17}} - 0.38. \quad (4.12)$$

Note that (4.11) contains a correction term m_{max} : The mass, M_{ecl} , between m_L and m_{max} does not include the star with m_{max} as this star lies between m_{max} and m_{max^*} . The semi-analytical calculation of the $m_{\text{max}} - M_{\text{ecl}}$ relation by Kroupa and Weidner (2003) and

Weidner et al. (2010) is less accurate by not including the correction term m_{\max} . The correction turns out to be insignificant as both semi-analytical relations are next to identical (► Fig. 4-2) and are a surprisingly good description of the observational data (► Fig. 4-5).

► Equation 4.9 defines, here for the first time, how to sample the IMF perfectly in the sense that the stellar masses are spaced ideally such that no gaps arise and the whole accessible range m_L to m_{\max} is fully filled with stars. This is referred to as optimal sampling (see also equation 8.2 in Aarseth 2003 for a related concept). The disadvantage of this method is that the target mass M_{ecl} cannot be achieved exactly because it needs to be distributed into a discrete number of stars. The mass of the generated stellar population is M_{ecl} to within $\pm m_L$ ($\equiv m_{\text{low}} \approx 0.07 M_{\odot}$ for most applications) because the integral (► 4.9) is integrated from m_{\max} downward.⁸

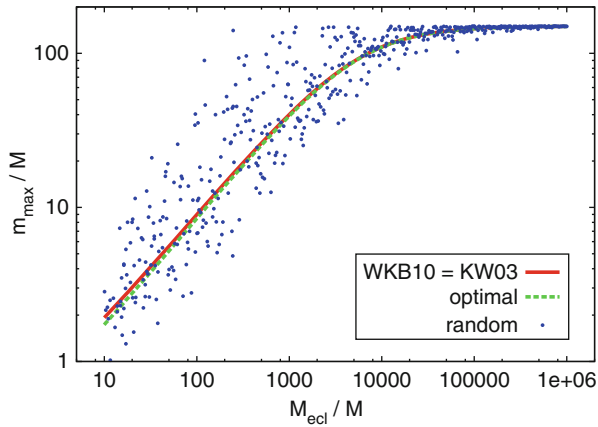
► Figure 4-1 demonstrates how an IMF constructed using optimal sampling compares to one generated with random sampling from the IMF for a population with $M_{\text{ecl}} = 150 M_{\odot}$. ► Figures 4-2 and ► 4-3, respectively, show $m_{\max} - M_{\text{ecl}}$ and average stellar mass model data using both sampling techniques.



■ Fig. 4-1

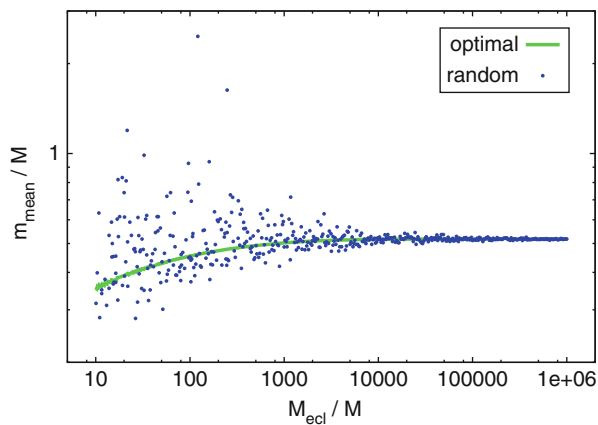
Stellar IMF (dN/dm) versus stellar mass (m) for a $M_{\text{ecl}} = 149.97 M_{\odot}$ cluster. The *thin dashed line* is the analytical canonical IMF (► 4.55). The *thick histogram* is a population of individual stars generated according to optimal sampling (► 4.9), starting with $m_{\max} = 12 M_{\odot}$ (► 4.12). The optimally sampled population contains 322 stars. The mass-dependent bin width is chosen to ensure that ten stars are in each bin. The *thin dashed histogram* is a population of stars chosen randomly from the IMF using the mass-generating function (► Sect. 2.3) with $M_{\text{ecl}} = 150.22 M_{\odot}$ being reached without an upper limit ($m_{\max} = \infty$). This population contains 140 stars. The same bins are used as in the *thick solid histogram*. Note how optimal sampling perfectly reproduces the IMF while random sampling shows deviations in the form of gaps. Which is closer to reality (remembering that observational data contain uncertainties that act as randomization agents)?

⁸The publicly available C program *Optimf* allowing the generation of a stellar population of mass M_{ecl} is available at <http://www.astro.uni-bonn.de/en/download/software/>



■ Fig. 4-2

The maximum stellar mass, m_{\max} , is plotted against the stellar mass of the population, M_{ecl} , for optimal sampling (*short-dashed green line*) and for randomly drawing a population of stars from the IMF (*filled blue circles*). For each sampling method, 100 populations are generated per dex in M_{ecl} . The canonical physical maximum stellar mass, $m_{\max*} = 150 M_{\odot}$, is assumed. The semi-analytical $m_{\max} - M_{\text{ecl}}$ relation from Kroupa and Weidner (2003) and Weidner et al. (2010) (*red solid curve*) and the corrected version from (4.12) (*green dashed curve*) are nearly identical. Which is closer to reality, random or optimal sampling from the IMF? The small scatter in the observational $m_{\max} - M_{\text{ecl}}$ relation (4.5) and the small scatter in observationally derived IMF power-law indices around the Salpeter value (4.27, Open Question III on p. 195) suggest that optimal sampling may be a more realistic approach to nature than purely random sampling



■ Fig. 4-3

The average stellar mass of stellar populations of mass M_{ecl} generated with optimal sampling and random sampling from the IMF; otherwise as (4.2) Note the extreme random deviations from the canonical IMF if it is interpreted to be a probabilistic distribution function

Important Hint

Nature may be preferring optimal sampling. This is evident from the quite tight observational $m_{\max} - M_{\text{ecl}}$ data (► Fig. 4-5), from Open Question III (p. 195) and from the Sociological Hypothesis (p. 196). Theoretical support that nature may not be playing dice comes from the emergence of the $m_{\max} - M_{\text{ecl}}$ relation in SPH and FLASH computations of star formation in turbulent molecular clouds (see the BVB conjecture on p. 148).

2.3 Discretizing an IMF: Random Sampling and the Mass-Generating Function

In order to randomly discretize a stellar population, we need to be able to generate stellar masses independently of each other. This can be done with constraints (e.g., ensuring that the $m_{\max} - M_{\text{ecl}}$ relation is fulfilled by applying $m \leq m_{\max}(M_{\text{ecl}})$ or that stellar masses $m \leq m_{\max*} \approx 150 M_{\odot}$) or without constraints ($m_{\max} = \infty$). This is achieved via the mass-generating function. A *mass-generating function* is a mapping from a uniformly distributed random variable X to the stellar mass which is distributed according to the IMF. Generating functions allow efficient random discretization of continuous distributions (see Kroupa 2008 for more details).

A generating function can be written in the following way. Assume the probability distribution function depends on the variable $\zeta_{\min} \leq \zeta \leq \zeta_{\max}$ (in this case the stellar mass, m). Consider the probability, $X(\zeta)$, of encountering a value for the variable in the range ζ_{\min} to ζ ,

$$X(\zeta) = \int_{\zeta_{\min}}^{\zeta} p(\zeta') d\zeta', \quad (4.13)$$

with $X(\zeta_{\min}) = 0 \leq X(\zeta) \leq X(\zeta_{\max}) = 1$ and $p(\zeta)$ is the probability distribution function, or probability density, normalized such that the latter equal sign holds ($X = 1$). For the two-part power-law IMF, the corresponding probability density is

$$\begin{aligned} p_1(m) &= k_{p,1} m^{-\alpha_1}, \quad 0.07 \leq m \leq 0.5 M_{\odot} \\ p_2(m) &= k_{p,2} m^{-\alpha_2}, \quad 0.5 \leq m \leq m_{\max}, \end{aligned} \quad (4.14)$$

where $k_{p,i}$ are normalization constants ensuring continuity at $0.5 M_{\odot}$ and

$$\int_{0.07 M_{\odot}}^{0.5 M_{\odot}} p_1 dm' + \int_{0.5 M_{\odot}}^{m_{\max}} p_2 dm' = 1, \quad (4.15)$$

whereby m_{\max} may follow from the mass of the population, M_{ecl} . Defining

$$X'_1 = \int_{0.07 M_{\odot}}^{0.5 M_{\odot}} p_1(m') dm', \quad (4.16)$$

it follows that

$$X_1(m) = \int_{0.07 M_{\odot}}^m p_1(m') dm', \quad \text{if } m \leq 0.5 M_{\odot}, \quad (4.17)$$

or

$$X_2(m) = X'_1 + \int_{0.5 M_{\odot}}^m p_2(m') dm', \quad \text{if } m > 0.5 M_{\odot}. \quad (4.18)$$

The generating function for stellar masses follows by inverting the above two equations $X_i(m)$.

The procedure is then to choose a uniformly distributed random variate $X \in [0, 1]$ and to select the generating function $m(X_1 = X)$ if $0 \leq X \leq X'_1$, or $m(X_2 = X)$ if $X'_1 \leq X \leq 1$. Stellar masses are generated until M_{ecl} is reached to within some preset tolerance. This algorithm is readily generalized to any number of power-law segments (([Sect. 4.5](#)), ([Sect. 9.1](#)), such as including a third segment for brown dwarfs and allowing the IMF to be discontinuous near $0.07 M_\odot$ ([Sect. 8.4](#)). Such a form has been incorporated into Aarseth's N -body4/6/7 programs.

For a general $\xi(m)$ and if $X(m)$ cannot be inverted, stellar masses may be generated by constructing a table of $X(m)$, m values,

$$M(m) = \int_{0.07 M_\odot}^m m' \xi(m') dm', \quad X(m) = \frac{M(m)}{M_{\text{ecl}}}, \quad (4.19)$$

such that $X(m_{\text{max}}) = 1$. For a random variate X , the corresponding m is obtained by interpolating the table, whereby the X is distributed uniformly and the procedure is repeated until M_{ecl} is reached to some preset tolerance.

Another highly efficient method for generating stellar masses randomly from arbitrary distribution functions is discussed in ([Sect. 2.4](#)).

2.4 A Practical Numerical Formulation of the IMF

Assuming the stellar IMF to be a probability density distribution such that stellar masses can be generated randomly ([Sect. 2.3](#)) from the IMF with ($m \leq m_{\text{max}}$) or without ($m_{\text{max}} = \infty$) constraints remains a popular approach. The two-part description can be straightforwardly expanded to a multipart power law. However, the direct implementation of this description requires multiple IF statements. In the following, a handy numerical formulation is presented for randomly generating stellar masses which replaces complicated IF constructions by two straightforward loops (Pflamm-Altenburg and Kroupa 2006).

Historically, multi-power-law IMFs start indexing intervals and slopes at zero instead of one. For simplicity, we here index n intervals from 1 up to n . We now consider an arbitrary IMF with n intervals fixed by the mass array $[m_0, \dots, m_n]$ and the array of functions f_1, \dots, f_n . On the i -th interval $[m_{i-1}, m_i]$, the IMF is described by the function f_i . The segment functions refer to the "linear" IMF, $\xi(m) = dN/dm$, and not to the logarithmic IMF, $\xi_L(m) = dN/d \log_{10} m$. At this point, it is not required that the segment functions f_i are scaled by a constant such that continuity is ensured on the interval boundaries. They only need to describe the functional form.

For the case of a multi-power law, these functions are

$$f_i(m) = m^{-\alpha_i}. \quad (4.20)$$

The segment functions may also be log-normal distributions, as, for example, in the IMFs of Miller and Scalo (1979) or Chabrier (2003a), but in general, they can be arbitrary.

We first define the two Θ -mappings (Θ -closed and Θ -open)

$$\Theta_{[]}(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}, \quad (4.21)$$

$$\Theta_{] [}(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}, \quad (4.22)$$

and the function

$$\Gamma_{[i]}(m) = \Theta_{[]}(m - m_{i-1})\Theta_{[]}(m_i - m) . \quad (4.23)$$

The $\Gamma_{[i]}(m)$ function is unity on the interval $[m_{i-1}, m_i]$ and zero otherwise.

The complete IMF can now be conveniently formulated by

$$\xi(m) = k \prod_{j=1}^{n-1} \Delta(m - m_j) \sum_{i=1}^n \Gamma_{[i]}(m) C_i f_i(m) , \quad (4.24)$$

where k is a normalization constant and the array (C_1, \dots, C_n) is to ensure continuity at the interval boundaries. They are defined recursively by

$$C_1 = 1 \quad , \quad C_i = C_{i-1} \frac{f_{i-1}(m_{i-1})}{f_i(m_{i-1})} . \quad (4.25)$$

For a given mass m , the $\Gamma_{[i]}$ makes all summands zero except the one in which m lies. Only on the inner interval boundaries do both adjoined intervals give the same contribution to the total value. The product over

$$\Delta(x) = \begin{cases} 0.5 & x = 0 \\ 1 & x \neq 0 \end{cases} \quad (4.26)$$

halves the value due to this double counting at the interval boundaries. In the case of n equals one (one single power law), the empty product has, by convention, the value of unity.

An arbitrary integral over the IMF is evaluated by

$$\int_a^b \xi(m) dm = \int_{m_0}^b \xi(m) dm - \int_{m_0}^a \xi(m) dm , \quad (4.27)$$

where the primitive of the IMF is given by

$$\begin{aligned} \int_{m_0}^a \xi(m) dm &= k \sum_{i=1}^n \Theta_{[]}(a - m_i) C_i \int_{m_{i-1}}^{m_i} f_i(m) dm \\ &+ k \sum_{i=1}^n \Gamma_{[i]}(a) C_i \int_{m_{i-1}}^a f_i(m) dm . \end{aligned} \quad (4.28)$$

The expressions for the mass content, that is, $m \xi(m)$, and its primitive are obtained by multiplying the above expressions in the integrals by m and one has to find the primitives of $m f_i(m)$.

Stars can now be diced from an IMF, $\xi(m)$, based on the above formulation and the concept of the generating function (Sect. 2.3) in the following way: A random number X is drawn from a uniform distribution and then transformed into a mass m . The mass segments transformed into the X -space are fixed by the array $\lambda_0, \dots, \lambda_n$ defined by

$$\lambda_i = \int_{m_0}^{m_i} \xi(m) dm . \quad (4.29)$$

If $P(X)$ denotes the uniform distribution with $P(X) = 1$ between 0 and λ_n , both functions are related by

$$\int_{m_0}^{m(X)} \xi(m') dm' = \int_0^X P(X') dX' = X . \quad (4.30)$$

If a given X lies between λ_{i-1} and λ_i , the corresponding mass m lies in the i -th interval $[m_{i-1}, m_i]$ and it follows

$$X(m) = \lambda_{i-1} + k C_i (F_i(m) - F_i(m_{i-1})), \quad (4.31)$$

or

$$m(X) = F_i^{-1} \left(\frac{X - \lambda_{i-1}}{k C_i} + F_i(m_{i-1}) \right), \quad (4.32)$$

where F_i is a primitive of f_i and F_i^{-1} is the primitive's inverse mapping. The complete expression for the solution for m is given by

$$m(X) = \sum_{i=1}^n \lambda \Gamma_{[i]} F_i^{-1} \left(\frac{X - \lambda_{i-1}}{k C_i} + F_i(m_{i-1}) \right) \cdot \prod_{j=1}^{n-1} \Delta(X - \lambda_j), \quad (4.33)$$

where $\lambda \Gamma_i$ are mappings which are unity between λ_{i-1} and λ_i and zero otherwise. Note that the primitives are determined except for an additive constant, but it is canceled out in the relevant expressions in (4.28) and (4.33).

The most used segment function for the IMF is a power law,

$$f(m) = m^{-\alpha}. \quad (4.34)$$

The corresponding primitive and its inverse mapping is

$$F(m) = \begin{cases} \frac{m^{1-\alpha}}{1-\alpha} & \alpha \neq 1 \\ \ln(m) & \alpha = 1 \end{cases}, \quad (4.35)$$

and

$$F^{-1}(X) = \begin{cases} ((1-\alpha)X)^{\frac{1}{1-\alpha}} & \alpha \neq 1 \\ \exp(X) & \alpha = 1 \end{cases}. \quad (4.36)$$

The other segment function used is a log-normal distribution, that is, a Gaussian distribution of the logarithmic mass,

$$\xi(lm) \propto \exp \left(-\frac{(lm - lm_c)^2}{2\sigma^2} \right), \quad (4.37)$$

where $lm \equiv \log_{10} m$. The corresponding segment function is

$$f(m) = \frac{1}{m} \exp \left(-\frac{(lm - lm_c)^2}{2\sigma^2} \right), \quad (4.38)$$

with the primitive

$$F(m) = \sqrt{\frac{\pi}{2}} \sigma \ln 10 \operatorname{erf} \left(\frac{lm - lm_c}{\sqrt{2}\sigma} \right), \quad (4.39)$$

and the inverse of the primitive

$$F^{-1}(X) = 10^{\sqrt{2}\sigma \operatorname{erf}^{-1} \left(\sqrt{\frac{2}{\pi}} \frac{X}{\sigma \ln 10} \right) + lm_c}, \quad (4.40)$$

where erf and erf^{-1} are the Gaussian error function,

$$\text{erf}(Y) = \frac{2}{\sqrt{\pi}} \int_0^Y e^{-y^2} dy, \quad (4.41)$$

and its inverse, respectively.

Several accurate numerical approximations of the Gaussian error function exist, but approximations of its inverse are quite rare. One such handy numerical approximation of the Gaussian error function which allows an approximation of its inverse, too, has been presented by Sergei Winitzki⁹ based on a method explained in Winitzki (2003):

The approximation of the error function for $Y \geq 0$ is

$$\text{erf}(Y) \approx \left(1 - \exp\left(-Y^2 \frac{\frac{4}{\pi} + a Y^2}{1 + a Y^2}\right) \right)^{\frac{1}{2}}, \quad (4.42)$$

with

$$a = \frac{8}{3\pi} \frac{\pi - 3}{4 - \pi}. \quad (4.43)$$

Values for negative Y can be calculated with

$$\text{erf}(Y) = -\text{erf}(-Y). \quad (4.44)$$

The approximation for the inverse of the error function follows directly,

$$\text{erf}^{-1}(Y) \approx \left(-\frac{2}{\pi a} - \frac{\ln(1 - Y^2)}{2} + \sqrt{\left(\frac{2}{\pi a} + \frac{\ln(1 - Y^2)}{2} \right)^2 - \frac{1}{a} \ln(1 - Y^2)} \right)^{\frac{1}{2}}. \quad (4.45)$$

Important Result

The above algorithm for dicing stars from an IMF supporting power-law and log-normal segment functions has been coded in the publicly available software package libimf available at <http://www.astro.uni-bonn.de/download/software/>

2.5 Statistical Treatment of the Data

Whichever is a better description of nature, optimal or random sampling from the IMF, a set of observationally derived stellar masses will appear randomized because of uncorrelated measurement uncertainties. Statistical tools are therefore required to help analyze the observed set of masses in the context of their possible parent distribution function and upper limit.

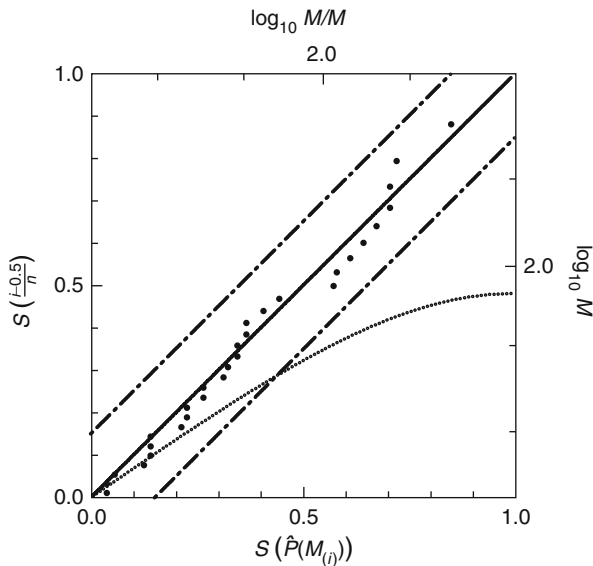
We concentrate here in particular on analyzing the high-mass end of the IMF, which follows a power-law probability density. Estimating the exponent via binning (using constant-size bins in logarithmic space) can introduce significant bias (Maíz Apellániz and Úbeda 2005; Maschberger and Kroupa 2009), especially for meager data sets. This can be remedied by using bins containing approximately the same number of data points (Maíz Apellániz and Úbeda 2005).

⁹homepages.physik.uni-muenchen.de/~Winitzki/erf-approx.pdf

However, binning does not allow one to estimate the upper limit of the mass function. A more suitable approach to estimate both exponent and upper limit simultaneously is to use the maximum likelihood method. The estimate there is given just by the largest data point and is consequently also naturally biased to too small values. Maschberger and Kroupa (2009) give a correction factor which leads to unbiased results for the upper limit.


Besides biases due to the statistical method, observational limitations can also introduce biases in the exponent. The influence of unresolved binaries for the high-mass IMF slope is less than ± 0.1 dex (Maíz Apellániz 2008; Weidner et al. 2009). Random superpositions, in contrast, can cause significant biases, as found by Maíz Apellániz (2008). Differential reddening can affect the deduced shape of the IMF significantly (Andersen et al. 2009). A further point in data analysis besides the estimation of the parameters is to validate the assumed power-law form for the IMF and in particular to decide whether, for example, a universal upper limit ($m_{\max^*} = 150M_{\odot}$) is in agreement with the data.

For this purpose, standard statistical tests, for example, the Kolmogorov–Smirnov test, can be utilized, but their deciding powers are not very high. They can be significantly improved by making a stabilizing transformation, S (Maschberger and Kroupa 2009). A graphical goodness-of-fit assessment can then be made using the stabilized probability-probability (SPP) plot, for example, \blacktriangleright Fig. 4-4. This plot has been constructed using a truncated power law ($m_{\max^*} = 143M_{\odot}$) and is aimed to help to decide whether a truncated or an infinite power law fits the stellar masses of the 29 most massive stars in R136 (masses taken from Massey and Hunter



\blacksquare Fig. 4-4

SPP plot of the massive stars in R136 using a power law truncated at $143M_{\odot}$ (From Maschberger and Kroupa 2009). The data are following this hypothesis and lie along the diagonal within the 95% acceptance region of the stabilized Kolmogorov–Smirnov test (limited by *dashed lines*). An infinite power law as the parent distribution function, the *dotted line* bending away from the diagonal, can be ruled out. The stellar MF in R136 is thus a power law with index $\alpha = 2.2$ truncated at $143M_{\odot}$


1998, using the isochrones of Chlebowski and Garmany 1991). The points are constructed from the ordered sample of the masses, with the value of the stabilized cumulative probability, $S(\hat{P}(m_{(i)}))$ as x-coordinate (using the estimated parameters) and the stabilized empirical cumulative probability, $S(\frac{i-0.5}{n})$, as the y-coordinate. The data follow the null hypothesis of a truncated power law and lie therefore along the diagonal. The null hypothesis would be challenged significantly if the data would leave the region enclosed by the dashed parallels to the diagonal, which is the 95% acceptance region of the stabilized Kolmogorov–Smirnov test. An alternative hypothesis of an infinite power law (with the same exponent) is shown as the dotted line which strongly bends away from the diagonal and is in significant disagreement with the data. Plots like  Fig. 4-4 can be constructed for any combination of null hypothesis and alternative hypothesis and have great potential to improve the statistical analysis of the upper mass end of the IMF.

Important Result

A program to estimate the parameter of power-law distributed data and to calculate goodness-of-fit tests for them is available in the publicly available software package statpl available at <http://www.astro.uni-bonn.de/download/software/>

2.6 Binary Systems

In order to infer the stellar IMF, it is necessary to account for all stars in the population, including the companions of multiple systems. Indeed, the vast majority of stars are observed to form in multiple systems. However, since dynamically not evolved star-forming regions of an age around about 1 Myr such as the Taurus-Auriga subclusters have a high multiplicity fraction of nearly 100%, this immediately implies that most stars by far must form as binaries. This is because if they were to form as triple or higher-order systems, then they would decay on their system crossing time scale which is far shorter than 1 Myr leading, by 0.5–1 Myr, to a substantial population of single stars which is not observed (Goodwin and Kroupa 2005; Goodwin et al. 2007).

The IMF appearing largely invariant to star-formation conditions (but see  Sect. 12.9) constitutes a statistical statement on one of the birth dynamical properties of stars (namely, their distribution of masses). So, since both the IMF and the birth binary population (BBP) are the result of the same (star formation) process and since the IMF is a result of this process “one level deeper down” than the BBP, it is quite natural to suggest that the formal mathematical distribution function of all of the birth dynamical properties (see footnote 4 on p. 121) of stars are also quite invariant. It follows that the star-formation outcome in terms of stellar masses (the IMF) and multiple systems (the birth binary population – BBP) can be formulated by the Star Formation Universality Hypothesis:

The Star-Formation Universality Hypothesis

IMF universality \iff BBP universality.

For stars with $m \lesssim 5 M_\odot$, the birth binary population is deduced from an elaborate analysis.

The birth binary population (BBP)

- Random pairing from the canonical IMF (► 4.55) for $0.1 \lesssim m/M_\odot \lesssim 5$.
- Thermal eccentricity distribution of eccentricities, $f_e(e) = 2e$.
- The period distribution function

$$f_{p,\text{birth}} = \eta \frac{IP - IP_{\min}}{\delta + (IP - IP_{\min})^2}, \quad (4.46)$$

where $\eta = 2.5$, $\delta = 45$, $IP_{\min} = 1$, and $\int_{IP_{\min}}^{IP_{\max}} f_{p,\text{birth}} dIP = 1$ such that the birth binary fraction is unity ($IP_{\max} = 8.43$; $IP \equiv \log_{10} P$ and P is in days). Here, $f_e de$ and $f_p dIP$ are the fraction of all orbits with eccentricity in the range $e, e + de$ and log period in the range $IP, IP + dIP$, respectively. Note that conversion between P and the semimajor axis is convenient through Kepler's third law: $a^3/P_{\text{year}}^2 = m_1 + m_2$, where a is in AU, $P_{\text{year}} = P/365.25$ is the period in years, and m_1, m_2 are the primary- and secondary-star masses in M_\odot .

The following are to be noted:

1. The BBP was derived by using observations of pre-main sequence and main-sequence binary populations as initial and final boundary conditions, respectively. Kroupa (1995a) postulated that there exist two stellar-dynamical operators, Ω_p and Ω_q , which independently transform the period- and mass-ratio distribution functions (that are independent products of the star-formation process for the majority of binaries) between the initial and final states. It is possible to demonstrate that both Ω_p and Ω_q exist. Furthermore, they are equal and are given by a characteristic star cluster consisting of 200 binary systems with a characteristic radius of about 0.8 pc (Kroupa 1995c; Marks et al. 2011),

$$\Omega_p = \Omega_q \equiv \Omega = (200 \text{ binaries}, R_{\text{ch}} \approx 0.8 \text{ pc}). \quad (4.47)$$

Important Result

The interpretation of this result is that the BBP is, like the IMF, a fundamental outcome of star formation and that most stars in the MW disk stem from star-formation events that are dynamically equivalent to the characteristic, or dominant-mode, cluster.

2. The deduced maximum binary period in the BBP of $10^{8.43}$ d corresponds to a spatial scale of $\approx 10^4$ AU which is the typical dimension of a pre-stellar cloud core (Kirk et al. 2005).
3. The evolved BBP, which matches the Galactic field stellar and binary population, also accounts simultaneously for the individual-star and the system LFs (► Fig. 4-15).

The BBP is the deduced (Kroupa 1995a, c) outcome of star formation in low to intermediate density ($\rho \lesssim 10^5 M_\odot/\text{pc}^3$) cloud regions (e.g., embedded clusters), but the rules layed out in (► 4.46) may well be formally applicable to higher density regions as well, whereby wide binaries are naturally truncated due to close-packing.

Important Note

In this sense, the BBP is a formal mathematical description of the outcome of star formation. Just like the formal stellar IMF (► 4.55), it may never be accessible to observations (the IMF Unmeasurability Theorem, p. 129). But, just like the stellar IMF, it is extractable from the observations.

The binary fraction

$$f = \frac{N_{\text{bin}}}{N_{\text{bin}} + N_{\text{sing}}}, \quad (4.48)$$

where $N_{\text{bin}}, N_{\text{sing}}$ are the number of binary- and single-stellar systems in the survey, respectively, is high ($f_{\text{bin}} > 0.8$) in dynamically unevolved populations, whereas $f_{\text{bin}} \approx 0.5$ for typically open clusters and the Galactic field as follows from applying Ω on the BBP (i.e., by performing N -body integrations of dissolving star clusters, Kroupa 1995b; Marks et al. 2011).

Note that the BBP needs to be transformed to the *initial binary population* by the process of pre-main sequence eigenevolution (Kroupa 1995d), which introduces the observed correlations between mass ratio, eccentricity, and period for short-period binaries, while the dynamically evolved initial binary population yields the observed mass-ratio and period distribution functions with $f \approx 0.5$ (Marks and Kroupa 2011).

For $m \gtrsim 5 M_{\odot}$ stars, the pairing rules change perhaps, reflecting the outcome of star formation in dense regions such as in the cores of embedded clusters ($\rho \gtrsim 10^5 M_{\odot}/\text{pc}^3$). Using a large sample of young clusters (for a review, see Sana and Evans 2010 and also Sana et al. 2011a), it is found that at least 45–55% of O stars are spectroscopic binaries: The mass ratios for these are larger in comparison with the late-type stars above: Massive binaries have a flat mass-ratio distribution and $0.2 \leq q \leq 1$. These systems have short periods, typically less than about 10 d, but extend from 0.3 to $10^{3.5}$ d. The measured distribution function is provided by equation (5.2) in Sana and Evans (2010). The overall binary fraction among O stars is at least 85% (García and Mermilliod 2001) as the spectroscopic fraction is augmented by wider visual binaries with separations between 40 and 200 AU (Sana et al. 2011b). The vast spectroscopic survey by Chini et al. (2012) of about 800 O and B-type stars affirms such results and establishes, even for runaway stars, the very high binary fraction and q about 1 pairing.

This leads to the following question.

Open Question I

Why do the differing BBP properties between $0.1 M_{\odot}$ and a few M_{\odot} on the one hand side and above a few M_{\odot} on the other hand side not correspond to the structure evident in the IMF, which is a featureless power law above about $0.5 M_{\odot}$ with a flattening below this mass? (► Sect. 9.1).

Below about $0.1 M_{\odot}$ very low mass stars and brown dwarfs, with $f_{\text{bin,BD}} \approx 0.15 - 0.2$, follow entirely separate rules (► Sect. 8.1) being an accompanying but distinct population to stars.

It has so far not been possible to predict nor to fully understand the distribution of binary-star birth properties from theory. The currently most advanced self-consistent gravohydrodynamical simulation without feedback of star formation using the SPH technique (Moeckel and Bate 2010) leads to too compact clusters of about a 1,000 stars and brown dwarfs from which a binary population emerges which does not quite have the observed distribution

of periods and mass ratios. However, this may be due to the currently unavoidable omission of feedback which would limit the depth of the gravitational collapse perhaps alleviating the binary-star problem (Kroupa 2011).

3 The Maximum Stellar Mass

While the stellar IMF appears to have a universal two-part power-law form ((☛ 4.55) below), the existence of a physical truncation mass as a function of embedded star cluster mass would suggest a form of IMF variation (☛ Sect. 12.1). Here, the evidence for such a truncation is presented.

3.1 On the Existence of a Maximum Stellar Mass

The empirically determined range of stellar masses poses important constraints on the physics of stellar formation, structure and stellar evolution, as well as on the feedback energy injected into a galaxy's atmosphere by a population of brand-new stars. The physical limit at low masses is now well established (☛ Sect. 8), and an upper mass limit appears to have been found recently.

A theoretical physical limitation to stellar masses has been known since many decades. Eddington (1926) calculated the limit which is required to balance radiation pressure and gravity, the *Eddington limit*: $L_{\text{Edd}}/L_{\odot} \approx 3.5 \times 10^4 m/M_{\odot}$. Hydrostatic equilibrium will fail if a star of a certain mass m has a luminosity that exceeds this limit, which is the case for $m \gtrsim 60 M_{\odot}$. It is not clear if stars above this limit cannot exist, as massive stars are not fully radiative but have convective cores. But more massive stars will lose material rapidly due to strong stellar winds. Schwarzschild and Härm (1959) inferred a limit of $\approx 60 M_{\odot}$ beyond which stars should be destroyed due to pulsations. But later studies suggested that these may be damped (Beech and Mitalas 1994). Stothers (1992) showed that the limit increases to $m_{\text{max}^*} \approx 120\text{--}150 M_{\odot}$ for more recent Rogers–Iglesia opacities and for metallicities $[\text{Fe}/\text{H}] \approx 0$. For $[\text{Fe}/\text{H}] \approx -1$, $m_{\text{max}^*} \approx 90 M_{\odot}$. A larger physical mass limit at higher metallicity comes about because the stellar core is more compact, the pulsations driven by the core having a smaller amplitude, and because the opacities near the stellar boundary can change by larger factors than for more metal-poor stars during the heating and cooling phases of the pulsations thus damping the oscillations. Larger physical mass limits are thus allowed to reach pulsational instability.

Related to the pulsational instability limit is the problem that radiation pressure also opposes accretion for protostars that are shining above the Eddington luminosity. Therefore, the question remains how stars more massive than $60 M_{\odot}$ may be formed. Stellar formation models lead to a mass limit near $40\text{--}100 M_{\odot}$ imposed by feedback on a spherical accretion envelope (Kahn 1974; Wolfire and Cassinelli 1987). Some observations suggest that stars may be accreting material in disks and not in spheres (e.g., Chini et al. 2004). The higher density of the disk material may be able to overcome the radiation at the equator of the protostar. But it is unclear if the accretion rate can be boosted above the mass-loss rate from stellar winds by this mechanism. Theoretical work on the formation of massive stars through disk accretion with high accretion rates thereby allowing thermal radiation to escape polewards (e.g., Jijina and Adams 1996) indeed lessen the problem and allow stars with larger masses to form.

Another solution proposed is the merging scenario. In this case, massive stars form through the merging of intermediate-mass protostars in the cores of dense stellar clusters driven by core contraction due to very rapid accretion of gas with low specific angular momentum, thus again avoiding the theoretical feedback-induced mass limit (Bonnell et al. 1998; Stahler et al. 2000, and the review by Zinnecker and Yorke 2007). It is unclear though if the very large central densities required for this process to act are achieved in reality, but it should be kept in mind that an observable young cluster is, by necessity, exposed from its natal cloud and is therefore likely to be always observed in an expanding phase such that the true maximally reached central density may be very high for massive clusters, $\approx 10^8 M_{\odot}/\text{pc}^3$ (Dabringhausen et al. 2010; Marks and Kroupa 2010; Conroy 2011).

The search for a possible maximal stellar mass can only be performed in massive, star-burst clusters that contain sufficiently many stars to sample the stellar IMF beyond $100 M_{\odot}$. Observationally, the existence of a finite physical stellar mass limit was not evident until very recently. Indeed, observations in the 1980s of R136 in the Large Magellanic Cloud (LMC) suggested this object to be one single star with a mass of about 2,000–3,000 M_{\odot} . Weigelt and Baier (1985) for the first time resolved the object into eight components using digital speckle interferometry, therewith proving that R136 is a massive star cluster rather than one single supermassive star. The evidence for any physical upper mass limit was very uncertain, and Elmegreen (1997) stated that “observational data on an upper mass cutoff are scarce, and it is not included in our models (of the IMF from random sampling in a turbulent fractal cloud).” Although Massey and Hunter (1998) found stars in R136 with masses ranging up to 140–155 M_{\odot} , Massey (2003) explained that the observed limitation is statistical rather than physical. We refer to this as the *Massey assertion*, that is, that $m_{\text{max}^*} = \infty$. Meanwhile, Selman et al. (1999) found, from their observations, a probable upper mass limit in the LMC near about 130 M_{\odot} , but they did not evaluate the statistical significance of this suggestion. Figer (2003) discussed the apparent cutoff of the stellar mass spectrum near 150 M_{\odot} in the Arches cluster near the Galactic center, but again did not attach a statistical analysis of the significance of this observation. Elmegreen (2000) also noted that random sampling from an unlimited IMF for all star-forming regions in the Milky Way (MW) would lead to the prediction of stars with masses $\gtrsim 1,000 M_{\odot}$, unless there is a rapid turndown in the IMF beyond several hundred M_{\odot} . However, he also stated that no upper mass limit to star formation has ever been observed, a view also emphasized by Larson (2003).

Thus, while theory clearly expected a physical stellar upper mass limit, the observational evidence in support of this was very unclear. This, however, changed in 2004.

3.2 The Upper Physical Stellar Mass Limit

Given the observed sharp drop-off of the IMF in R136 near 150 M_{\odot} , that is, that the R136 stellar population is observed to be saturated (p. 148), Weidner and Kroupa (2004) studied the above *Massey assertion* in some detail. R136 has an age ≤ 2.5 Myr (Massey and Hunter 1998) which is young enough such that stellar evolution will not have removed stars through supernova explosions. It has a metallicity of $[\text{Fe}/\text{H}] \approx -0.5$ dex (de Boer et al. 1985). From the radial surface density profile, Selman et al. (1999) estimated there to be 1,350 stars with masses between 10 and 40 M_{\odot} within 20 pc of the 30 Doradus region, within the center of which lies R136. Massey and Hunter (1998) and Selman et al. (1999) found that the IMF can be well approximated by a Salpeter power law with exponent $\alpha = 2.35$ for stars in the mass range 3–120 M_{\odot} .

(see also [Fig. 4-4](#)). This corresponds to 8,000 stars with a total mass of $0.68 \times 10^5 M_\odot$. Extrapolating down to $0.1 M_\odot$, the cluster would contain 8×10^5 stars with a total mass of $2.8 \times 10^5 M_\odot$. Using a canonical IMF with a slope of $\alpha = 1.3$ (instead of the Salpeter value of 2.35) between 0.1 and $0.5 M_\odot$, this would change to 3.4×10^5 stars with a combined mass of $2 \times 10^5 M_\odot$ for an average mass of $0.61 M_\odot$ over the mass range 0.1 – $120 M_\odot$. Based on the observations by Selman et al. (1999), Weidner and Kroupa (2004) assumed that R136 has a mass in the range $5 \times 10^4 \leq M_{\text{R136}}/M_\odot \leq 2.5 \times 10^5$. Using the *canonical stellar IMF* ([Eq. 4.55](#) below), they found that $N(> 150 M_\odot) = 40$ stars are missing if $M_{\text{ecl}} = 2.5 \times 10^5 M_\odot$, while $N(> 150 M_\odot) = 10$ stars are missing if $M_{\text{ecl}} = 5 \times 10^4 M_\odot$. The probability that no stars are observed although 10 are expected, assuming $m_{\text{max}*} = \infty$, is $P = 4.5 \times 10^{-5}$. Thus, the observations of the massive stellar content of R136 suggest a physical stellar mass limit near $m_{\text{max}*} = 150 M_\odot$.

A reanalysis of the stellar spectra plus new stellar modeling suggests, however, $m_{\text{max}*} \approx 300 M_\odot$ for R136 (Crowther et al. 2010). But Banerjee et al. (2012) demonstrate that $m > 150 M_\odot$ stars form readily from merging binaries in star-burst clusters. Their high-precision Aarseth- N -body models of binary-rich initially mass-segregated R136-type clusters demonstrate that stars much more massive than the $m_{\text{max}*} = 150 M_\odot$ limit appear from massive binaries that merge after becoming eccentric and hard through a stellar-dynamical encounter near the cluster core. Such binaries may be ejected from the cluster before merging, thus appearing to an observer as free-floating single stars of mass up to $300 M_\odot$. The fundamental upper mass limit may thus nevertheless be $m_{\text{max}*} \approx 150 M_\odot$.

Results similar to those of Weidner and Kroupa (2004) were obtained by Figer (2005) for the Arches cluster. The Arches is a star-burst cluster within 25 pc in projected distance from the Galactic center. It has a mass $M \approx 1 \times 10^5 M_\odot$ (Bosch et al. 2001), age 2–2.5 Myr and $[\text{Fe}/\text{H}] \approx 0$ (Najarro et al. 2004). It is thus a counterpart to R136 in that the Arches is metal rich and was born in a very different tidal environment to R136. Using his HST observations of the Arches, Figer (2005) performed the same analysis as Weidner and Kroupa (2004) did for R136. The Arches appears to be dynamically evolved, with substantial stellar loss through the strong tidal forces (Portegies Zwart et al. 2002), and the stellar mass function with $\alpha = 1.9$ is thus flatter than the Salpeter IMF. Using his updated IMF measurement, Figer calculated the expected number of stars above $150 M_\odot$ to be 33, while a Salpeter IMF would predict there to be 18 stars. Observing no stars but expecting to see 18 has a probability of $P = 10^{-8}$, again strongly suggesting $m_{\text{max}*} \approx 150 M_\odot$. The Arches cluster is thus another example of a saturated stellar population.

Given the importance of knowing if a finite physical upper mass limit exists and how it varies with metallicity, Oey and Clarke (2005) studied the massive-star content in nine clusters and OB associations in the MW, the LMC and the SMC. They predicted the expected masses of the most massive stars in these clusters for different upper mass limits (120, 150, 200, 1,000, and 10,000 M_\odot). For all populations, they found that the observed number of massive stars supports with high statistical significance the existence of a general upper mass cutoff in the range $m_{\text{max}*} \in (120, 200 M_\odot)$.¹⁰

The general indication thus is that a physical stellar mass limit near $150 M_\odot$ seems to exist. While biases due to unresolved multiples that may reduce the true maximal mass need to be studied further, the absence of variations of $m_{\text{max}*}$ with metallicity poses a problem.

¹⁰More recent work on the physical upper mass limit can be found in Koen (2006) and Maíz Apellániz et al. (2007, 2008).

A constant $m_{\max*}$ would only be apparent for a true variation as proposed by the theoretical models; *if metal-poor environments have a larger stellar multiplicity*, the effects of which would have to compensate the true increase of $m_{\max*}$ with metallicity. Interestingly, in a recent hydrodynamical calculation, Machida et al. (2009) find a higher binary fraction for low metallicities.

3.3 The Maximal Stellar Mass in a Cluster, Optimal Sampling and Saturated Populations

Above, we have seen that there seems to exist a universal physical stellar mass limit, $m_{\max*}$. However, an elementary argument suggests that star clusters must additionally limit the masses of their constituent stars: A pre-star-cluster gas core with a mass M_{core} can, obviously, not form stars with masses $m > \epsilon M_{\text{core}}$, where $\epsilon \approx 0.33$ is the star-formation efficiency (Lada and Lada 2003). Thus, given a freshly hatched cluster with stellar mass M_{ecl} , stars in that cluster cannot surpass masses $m_{\max} = M_{\text{ecl}}$, which is the identity relation corresponding to a “cluster” consisting of one massive star. Note that if we were to construct a (unphysical) model in which N stars are stochastically chosen from the IMF then such a constraint would not appear.

3.3.1 Theory

As discussed by Smith et al. (2009), there are two main theories of massive star formation: The first theory is essentially a scaled up version of low-mass star formation, where massive stars form from well-defined massive cores supported by turbulence. This model requires the existence of massive pre-stellar cores that manage to evade fragmentation during their formation stages. Perhaps radiative feedback can limit the fragmentation, but Smith et al. (2009) demonstrate that radiative feedback does not lead to the formation of massive pre-stellar cores in isolation.

The second theory is based on the *competitive accretion scenario* or on a refined version thereof, the *fragmentation limited starvation model*. Cores are the seeds of the formation of stars, and the most massive of these have a larger gravitational radius of influence and are therewith more successful at accreting additional mass. They can thus grow into massive stars. The massive seeds typically tend to form and stay at the center of the gravitational potential of the forming star cluster which they contribute to because the gas densities and thus the accretion rates are largest there. They accrete material via Bondi–Hoyle accretion, but when the velocity relative to the system is low, the accretion is mainly regulated by the tidal field. There is no requirement for stellar mergers, which, however can occur in dense regions (➤ 4.60) and contribute to the buildup of the IMF.

The first theory would imply that the formation of massive stars can occur in isolation, that is, without an accompanying star cluster. The existing data on the spatial distribution of massive stars do, however, not support this possibility (➤ Sect. 4). The second theory requires massive stars to be associated with star clusters. This is demonstrated by Smith et al. (2009) using SPH simulations of a gas cloud with a changing equation of state as a result of the cooling process shifting from line emission to dust emission with increasing density and with radiative heating as a model of the feedback process. Peters et al. (2010, 2011b) study this issue independently with three-dimensional, radiation-hydrodynamical simulations that include heating

by ionizing and non-ionizing radiation using the adaptive-mesh code FLASH and verify that massive star formation is associated with low-mass star formation as fragmentation cannot be suppressed, even including radiative feedback. Simpler isothermal SPH computations without feedback by Bonnell et al. (2004) and the FLASH simulations by Peters et al. (2010, 2011b) show that the most massive star in the forming cluster evolves with time, t , according to the following relation:

$$m_{\max}(t) = 0.39 M_{\text{ecl}}(t)^{2/3}. \quad (4.49)$$

The general form of the observed IMF (◆ 4.55) is also obtained. This is therefore a prediction of the second theory whereby it is important to note that (◆ 4.49) is a result of purely gravitationally driven star formation calculations with and without feedback. Interestingly, Peters et al. (2010, 2011b) conclude that computations with feedback lead to a closer agreement with the observed $m_{\max} - M_{\text{ecl}}$ data. The following result emerges.

The Bonnell-Vine-Bate (BVB) Conjecture

“Thus an individual cluster grows in numbers of stars as the most massive star increases in mass. This results in a direct correlation . . . , and provides a physical alternative to a probabilistic sampling from an IMF” (Bonnell et al. 2004).

Main result: The self-consistent gravo-hydrodynamical simulations of star formation thus yield the result that the growth of the most massive star is intimately connected with the growth of its hosting cluster, thereby populating the stellar IMF.

How does nature arrange the mass of the star-forming material over the emerging stellar masses? That massive stars can form in isolation can be discarded statistically (◆ Sect. 4). But how regulated or rather deterministic is the formation of massive stars and their star clusters?

Here, it is useful to return to the concept of optimal sampling (◆ Sect. 2.2): IF nature distributes the available mass M_{ecl} optimally over the IMF, then the $m_{\max} - M_{\text{ecl}}$ relation emerges. It is plotted in ◆ Fig. 4-5 as the thick-solid curves for two values of $m_{\max*}$.

For a given M_{ecl} , the observationally derived m_{\max} values show a spread rather than one value: Can stars with masses larger than the optimal m_{\max} , or even with masses beyond the canonical upper mass limit of $m_{\max*}$, occur? In discussing these issues, it proves useful to define the concept of a *saturated population*.

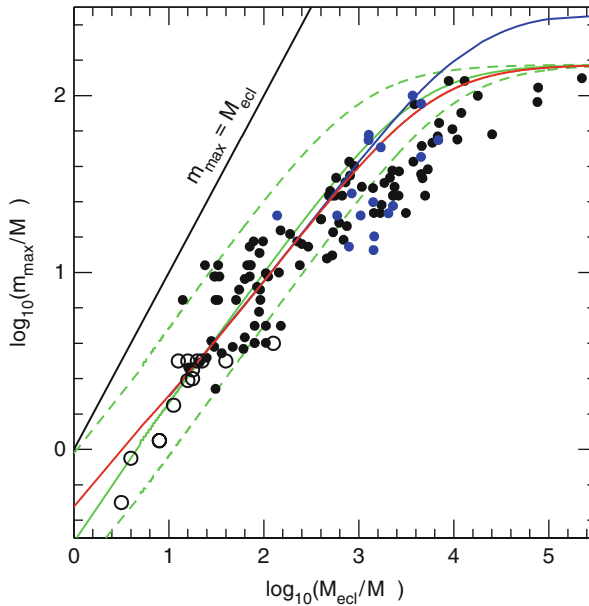
Definitions

Saturated population: A population for which $m_{\max} = m_{\max*}$ and $\int_{m_{\max}}^{\infty} \xi(m) dm \geq 1$. Only a simple population can be saturated.

Unsaturated population: A population for which $m_{\max} < m_{\max*}$.

Supersaturated population: A simple population containing super-canonical ($m > m_{\max*}$) stars.

Thus, a simple stellar population is saturated if its most-massive star has the physically allowed maximum mass (e.g., R136 and Arches). An unsaturated population is a simple population in which m_{\max} does not reach $m_{\max*}$ (e.g., ρ Oph and ONC). The star-burst cluster R136 is associated with super-canonical stars (Crowther et al. 2010) and is therefore a supersaturated population. This occurs naturally through merged massive binaries (Banerjee et al. 2012). Indeed, as an example that such stars are known to exist is discussed by Vanbeveren (2011).



■ Fig. 4-5

The $m_{\max} - M_{\text{ecl}}$ relation. The *black solid dots* are observed clusters (Weidner et al. 2010), with newly compiled data being shown as *blue solid dots*. The data are compiled subject to an age constraint (no cluster older than 4 Myr is accepted) and to the constraint that the cluster still be embedded. The *open circles* are the most-massive star versus cluster mass in 14 young stellar groups in Taurus, Lupus3, Chal, and IC348 (Kirk and Myers 2011). These are dynamically unevolved (Kroupa and Bouvier 2003b) and can be taken to represent pristine configurations. The *lower (red) thick solid line* is (4.12) with $m_{\max*} = 150 M_{\odot}$, and the *blue thick solid line* is the same but for $m_{\max*} = 300 M_{\odot}$. The *thin solid line* shows the identity relation, where a “cluster” consists only of one star. Assuming unconstrained random picking from the canonical IMF with $m_{\max*} = 150 M_{\odot}$, one sixth of all models would lie below the *lower (green) dashed curve*, while five sixth would lie below the *upper (green) dashed curve*, that is, two third of all data ought to lie between the *dashed curves*. The median, below which would lie 50% of all data for random sampling from the IMF, is plotted as the *thick (green) solid curve*. Significant deviations from random sampling from the stellar IMF are evident in that the scatter of the observational data is smaller and the median lies below the random-sampling median. The theoretical prediction (4.49) follows nearly precisely the *solid lines* for $M_{\text{ecl}} < 10^4 M_{\odot}$. Note also that the horizontal axis corresponds to a density. For example, if all clusters form with a half-mass radius of 0.5 pc, then the density scale becomes $\log_{10} \rho_* = \log_{10} M_{\text{ecl}}/M_{\odot} + 0.28$, where ρ_* is in units of M_{\odot}/pc^3 . Thus, stars more massive than $10 M_{\odot}$ appear only when $\rho_{\text{gas}} > 10^3 M_{\odot}/\text{pc}^3$ for a star-formation efficiency of 33%. Note that the clusters *saturate* (p. 148) for $M_{\text{ecl}} > 10^4 M_{\odot}$.

3.3.2 Observational data

A compilation of clusters only subject to a constraint in age and gas content (to ensure youth and dynamical virginity) for which the cluster mass and the initial mass of the heaviest star can be estimated observationally demonstrates that there exists a strong correlation between the

embedded-cluster mass and the most-massive stellar mass within it. The observational data are plotted in [Fig. 4-5](#).

By performing statistical tests, Weidner and Kroupa (2006) and Weidner et al. (2010) show with very high confidence that the observational data are not consistent with the most-massive star being randomly drawn from the IMF.¹¹ In addition, if the process of star formation was equivalent to pure random sampling of stars from the IMF, then this would predict the existence of star clusters dominated by O stars or even the formation of massive stars in isolation (e.g., Haas and Anders 2010). But this hypothesis is ruled out with high confidence because the theoretically *expected* fraction of massive stars that ought to appear to be isolated (1–4%) although they formed in clusters is already larger than the *observed* fraction of candidate isolated massive stars (see [Sect. 4](#)).

The newest additions of observational data enhance the empirical evidence for the existence of a physical $m_{\max} - M_{\text{ecl}}$ relation, even at masses $M_{\text{ecl}} \lesssim 15 M_{\odot}$: The low-mass data by Kirk and Myers (2011) show a remarkably small spread implying that even the lowest-mass “clusters” limit the mass of their most massive star in a nontrivial way. Indeed, the distribution of all data shows that random sampling from the IMF is ruled out since the spread is smaller than the expected spread given the 1/6th and 5/6th quantiles. The data in [Fig. 4-5](#) also show that for $M_{\text{ecl}} \lesssim 10^3 M_{\odot}$, the semi-analytical model ([4.12](#)) is an excellent description. At larger cluster masses, this model is a fairly good description as well although some systematic deviation is evident.

It thus appears that the process of star formation ends up close to optimally sampling the IMF and that it does not correspond to purely randomly generating stars from the IMF in support of the BVB conjecture (p. 148). But why?

Open Question II

Why is it that the star-formation process samples the IMF close to if not optimally?

3.3.3 Interpretation

The observational data suggest that the dominant physical process responsible for the $m_{\max} - M_{\text{ecl}}$ relation is a competitive resource-limited growth process. This would be natural since the protostars begin as a distribution of low-mass seed masses and accrete at various rates, thereby depleting the surrounding interstellar medium, as is in fact evident in self-consistent gravo-hydrodynamical simulations of star formation with and without feedback as discussed above.

For $M_{\text{ecl}} > 10^{2.2} M_{\odot}$, the small spread persists in the observational data, but the data fall below the semi-analytical model ([4.12](#)). This may hint at additional processes becoming important perhaps related to the ability for $m \gtrsim 10 M_{\odot}$ stars to continue to grow through

¹¹A study by Maschberger and Clarke (2008) of the most-massive star data in young star clusters concluded that “the data are not indicating any striking deviation from the expectations of random drawing.” This statement has been frequently misinterpreted that other sampling mechanisms are ruled out. However, the Maschberger and Clarke analysis focuses on low-mass clusters where the data were insufficient to decide whether star clusters are populated purely randomly from an IMF with constant upper mass limit or, for example, in a sorted fashion. The differences appear clearly at higher cluster masses, not included in their analysis but in Weidner and Kroupa (2006) and Weidner et al. (2010). Maschberger & Clarke (2008) adapt their data set according to the requested result and so their study does not constitute an acceptable scientific standard.

accretion alone. Another physical process of possible relevance leading to a reduction of $dm_{\max}/dM_{\text{ecl}}$ for $m_{\max} \gtrsim 10 M_{\odot}$ may be due to an instability in the cold inter-stellar medium (ISM) developing at a mass around $M_{\text{ecl}} = 10^2 M_{\odot}$ similar to the ISM instabilities discussed in Pflamm-Altenburg and Kroupa (2009a). Such an accretion instability may lead to enhanced accretion of gas onto the pre-cluster cloud core from the surrounding molecular cloud. If $m_{\max} \gtrsim 10 M_{\odot}$ stars have a reduced accretion efficiency, then this may explain the flattening since a smaller fraction of the newly accreted gas adds to the growth of m_{\max} and is instead used up by the formation of less-massive cluster stars (*fragmentation-induced starvation* of Peters et al. 2010, 2011b). Furthermore, it may be possible that for $M_{\text{ecl}} > 10^2 M_{\odot}$, sub-cluster merging may be becoming an important physical process: Each sub-cluster with $M_{\text{ecl}} \lesssim 100 M_{\odot}$ follows the $m_{\max} - M_{\text{ecl}}$ relation such that upon amalgamation of the sub-clusters, m_{\max} changes less than M_{ecl} . The steepening of the $m_{\max} - M_{\text{ecl}}$ relation for $M_{\text{ecl}} \gtrsim 10^3 M_{\odot}$ may be affected by the coalescence of massive protostars in the dense centers of forming embedded clusters.

Peters et al. (2010, 2011b) discuss the physics driving the $m_{\max} - M_{\text{ecl}}$ relation (for $M_{\text{ecl}} \lesssim 10^2 M_{\odot}$) and find that m_{\max} -growth curves flatten with increasing M_{ecl} because infalling gas is accreted by the other stars in the emerging cluster. In particular, the appearance of close companions to the most-massive star reduces its growth, while the star cluster continues to form. Feedback allows the growth of the most-massive star to be sustained for longer essentially by heating the gas such that it is less susceptible to fall into the potentials of lower-mass companions and stars and is therewith forced to follow the main potential toward the center, thereby leading to better agreement with the observed $m_{\max} - M_{\text{ecl}}$ relation. If no cluster of low-mass stars were to form such that none of the gas is accreted by the other low-mass stars, then $m_{\max} \propto M_{\text{ecl}}$, which is a dependency which is too steep compared to the data.

3.3.4 Stochastic or Regulated Star Formation?

The existence of an observed $m_{\max} \propto M_{\text{ecl}}^{2/3}$ relation (☉ 4.49) different to the one expected from random sampling from the IMF thus implies that the formation of massive stars is associated with surrounding low-mass star formation. This suggests that the formation of stars within the cloud cores is mostly governed by gravitationally driven growth processes in a medium with limited resources.

If the outcome of star formation were to be inherently stochastic, as is often assumed to be the case, in the sense that stars are randomly selected from the full IMF, then this would imply that stellar feedback would have to be, by stringent logical implication, the randomization agent. In other words, the well-ordered process of stars arising from a molecular cloud core by pure gravitationally driven accretion as shown to be the case by self-consistent gravohydrodynamical simulations would have to be upset completely through feedback processes. As there is no physically acceptable way that this might arise, and since in fact the work of Peters et al. (2010, 2011b) has demonstrated that feedback actually helps establish the $m_{\max} - M_{\text{ecl}}$ relation, it is concluded that star formation cannot be a random process, and its outcome cannot be described by randomly choosing stars from an IMF.¹²

¹²Choosing stars randomly from the IMF is, however, a good first approximation for many purposes of study.

Returning to [Sect. 1.5](#), the following two alternative hypotheses can thus be stated:

IMF Random Sampling Hypothesis

A star formation event always produces a probabilistically sampled IMF.

IMF Optimal Sampling Hypothesis

A star formation event always produces an optimally sampled IMF such that the $m_{\max} - M_{\text{ecl}}$ relation holds true.

As stated above, the IMF random sampling hypothesis can already be discarded on the basis of the existing data and simulations. But can the current data discard the IMF optimal sampling hypothesis? The scatter evident in [Fig. 4-5](#) may suggest that optimal sampling is ruled out, since if it were true for every cluster, then a one-to-one relation between m_{\max} and M_{ecl} would exist. However, the following effects contribute to introducing a dispersion of m_{\max} values for a given M_{ecl} , even if an exact $m_{\max} - M_{\text{ecl}}$ relation exists:

- The measurement of M_{ecl} and m_{\max} are very difficult and prone to uncertainty.
- An ensemble of embedded clusters of a given stellar mass M_{ecl} is likely to end up with a range of m_{\max} because the pre-cluster cloud cores are likely to be subject to different boundary conditions (internal distribution of angular momentum, different thermodynamic state, different external radiation field, different metallicity, etc.). The details of self-regulation depend on such quantities, and this may be compared to the natural dispersion of stellar luminosities at a given stellar mass due to different metallicity, stellar spins, and orientations relative to the observer and different stellar ages.

At present, the IMF optimal sampling hypothesis can thus not be discarded, but its statement here may be conducive to further research to investigate how exactly valid it is.

3.3.5 A Historical Note

Larson (1982) had pointed out that more massive and dense clouds correlate with the mass of the most massive stars within them, and he estimated that $m_{\max} = 0.33 M_{\text{cloud}}^{0.43}$ (masses are in M_{\odot}). An updated relation was derived by Larson (2003) by comparing m_{\max} with the stellar mass in a few clusters, $m_{\max} \approx 1.2 M_{\text{cluster}}^{0.45}$. Both are flatter than the semi-analytical relation and therefore do not fit the data in [Fig. 4-5](#) as well (Weidner and Kroupa 2006). Elmegreen (1983) constructed a relation between cluster mass and its most massive star based on an assumed equivalence between the luminosity of the cluster population and its binding energy for a Miller–Scalo IMF (a self-regulation model). This function is even shallower than the one estimated by Larson (2003).

3.4 Caveats

Unanswered questions regarding the formation and evolution of massive stars remain. There may be stars forming with $m > m_{\text{max}^*}$ which implode “invisibly” after 1 or 2 Myr. The explosion mechanism sensitively depends on the presently still rather uncertain mechanism for shock revival after core collapse (e.g., Janka 2001). Since such stars would not be apparent in massive clusters older than 2 Myr, they would not affect the empirical maximal stellar mass, and $m_{\text{max}^*,\text{true}}$ would be unknown at present.

Furthermore, stars are often in multiple systems. Especially, massive stars seem to have a binary fraction of 80% or even larger and apparently tend to be in binary systems with a preferred mass ratio $q \gtrsim 0.2$ (Sect. 2.6). Thus, if all O stars would be in equal-mass binaries, then $m_{\text{max}^*,\text{true}} \approx m_{\text{max}^*}/2$.

Finally, it is noteworthy that $m_{\text{max}^*} \approx 150 M_{\odot}$ appears to be the same for low-metallicity environments ($[\text{Fe}/\text{H}] = -0.5$, R136) and metal-rich environments ($[\text{Fe}/\text{H}] = 0$, Arches), in apparent contradiction to the theoretical values (Stothers 1992). Clearly, this issue needs further study.

Main Results

A fundamental upper mass limit for stars appears to exist, $m_{\text{max}^*} \approx 150 M_{\odot}$. The mass of a cluster defines the most-massive star in it and leads to the existence of a $m_{\text{max}} - M_{\text{ecl}}$ relation, which results from competitive resource-limited growth and self-regulation processes. The outcome of a star formation event appears to be close to an optimally sampled IMF.

4 The Isolated Formation of Massive Stars

An interesting problem relevant for the discussion in Sect. 3.3 with major implications for star-formation theory and the IMF in whole galaxies is whether massive stars can form alone without a star cluster, that is, in isolation (e.g., Li et al. 2003; Krumholz and McKee 2008).

Related to this is one of the most important issues in star-formation theory, namely, the still incomplete understanding of how massive stars ($m \gtrsim 10 M_{\odot}$) form. From Fig. 4-5, a gas density of $\rho_{\text{gas}} \gtrsim 10^3 M_{\odot}/\text{pc}^3$ for the formation of $m > 10 M_{\odot}$ stars can be deduced. At least four competing theories have been developed: the competitive accretion scenario (Bonnell et al. 1998, 2004; Bonnell and Bate 2006), collisional merging (Bonnell and Bate 2002), the single core collapse model allowing isolated massive star formation (Krumholz et al. 2009), the fragmentation-induced starvation model (Peters et al. 2010, 2011b) and the outflow-regulated clump-collapse models (Wang et al. 2010) where massive stars result from the collapse of turbulent cluster-forming clumps, whose internal dynamics are regulated by protostellar outflows.

To help advance this topic, it is necessary to find conclusive constraints for the formation of massive stars from observations. One important piece of evidence can be deduced from the formation sites of massive stars. While competitive accretion, collisional merging, and fragmentation-induced starvation descriptions and outflow-regulated clump collapse predict

the formation of massive stars within star clusters, the core collapse model needs a sufficiently massive and dense cloud core and allows for an isolated origin of O stars.

This latter model appears to be inconsistent with the data (► Sect. 3.3). And, even the currently most advanced radiation-magnetohydrodynamical simulations including ionization feedback of a $1,000 M_{\odot}$ rotating cloud lead to suppression of fragmentation by merely a factor of about two (Peters et al. 2011a), so that massive star formation cannot be separated from the formation of embedded clusters.

Nevertheless, the isolated formation of massive stars remains a popular option in the research community. Discussing field O stars, Massey (1998) writes (his p. 34) “One is tempted to conclude that these O3 stars formed pretty much where we see them today, as part of very modest star-formation events, ones that perhaps produce “a single O star plus some change,” as Jay Gallagher aptly put it.” Selier et al. (2011) write “There is, however, a statistically small percentage of massive stars ($\approx 5\%$) that form in isolation (de Wit et al. 2005; Parker and Goodwin 2007)” while Camargo et al. (2010) confess “On the other hand, de Wit et al. (2005) estimate that nearly 95% of the Galactic O star population is located in clusters or OB associations, or can be kinematically linked with them.” Lamb et al. (2010) state “In a study of Galactic field O stars, de Wit et al. (2004, 2005) find that 4 ± 2 per cent of all Galactic O stars appear to have formed in isolation, without the presence of a nearby cluster or evidence of a large space velocity indicative of a runaway star,” while Krumholz et al. (2010) explain “de Wit et al. (2004, 2005) find that 4 ± 2 per cent of galactic O stars formed outside of a cluster of significant mass, which is consistent with the models presented here (for example, runs M and H form effectively isolated massive single stars or binaries), but not with the proposed cluster-stellar mass correlation.” Such events of isolated massive star formation are conceivable if the equation of state of the interstellar medium can become stiff (Li et al. 2003). Hence, the search for isolated O stars and the deduction whether or not these stars have formed in situ can be vital in narrowing down theories and advancing the research field. Generally, in order to propose that massive stars form in isolation rather contrived initial conditions in the cloud core are required such as a strong magnetic field, no turbulence and a highly peaked density profile (Girichidis et al. 2011).

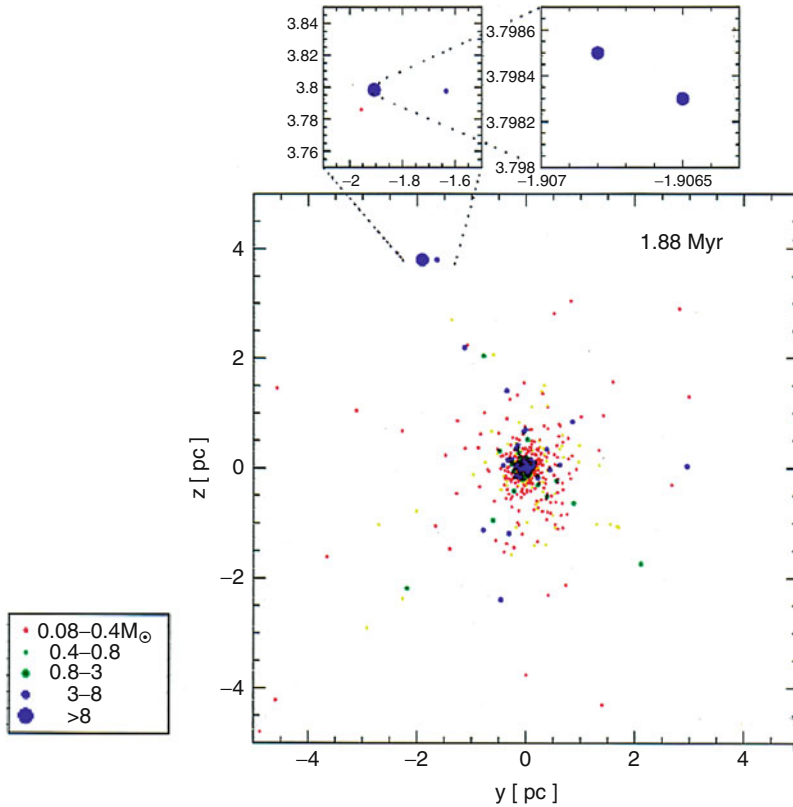
The existence of massive stars formed in isolation would be required if the stellar distribution within a galaxy was a result of a purely stochastic process in contradiction to the IGIMF theory (► Sect. 13.1), and would not be in agreement with the existence of the $m_{\max} - M_{\text{ec1}}$ relation (► 4.49).

This discussion does not have an easy observational solution because massive stars have been observed to be quite far away from sites where late-type stars or star clusters are forming and it can never be proven beyond any arbitrarily small doubt that a given star comes from a cluster.

One possible way to prove that a massive star formed in isolation, that is, with at most a few companions, would be to discover an isolated massive star with wide companions. This would invalidate the star having been ejected from a cluster. Unfortunately, even this possible criterion is not fool proof evidence for the occurrence of isolated massive star formation, as the example in ► Fig. 4-6 demonstrates.

Thus, a decision on whether star formation is random enough to allow the formation of massive stars by themselves in isolation can only be reached through statistical arguments since it can never be proven beyond doubt that some particular massive star did not form in isolation. It is thus essential to understand all possible physical mechanisms that contribute to massive stars being distributed widely throughout a galaxy.

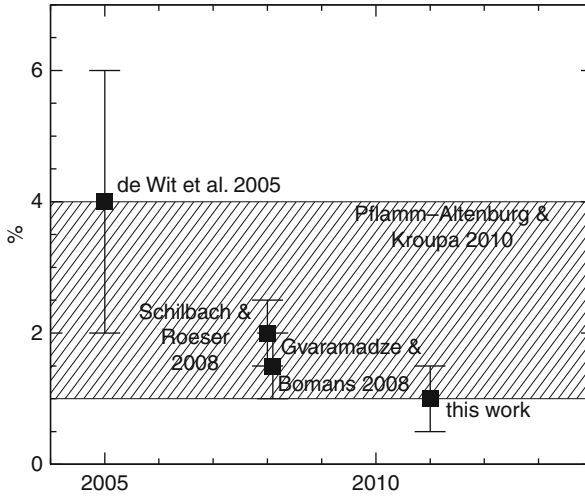
The “OB field star” MF has $\alpha \approx 4.5$, which has been interpreted to be the result of isolated high-mass star formation in small clouds (Massey 1998). Precise proper-motion measurements



■ Fig. 4-6

An N -body5 (Aarseth 1999) model of a star cluster initially not mass segregated, in virial equilibrium, with a randomly drawn IMF with stellar masses between 0.08 and $30 M_{\odot}$ and with a binary fraction of 100% (☉ 4.46). The cluster consists of 400 binaries distributed in a Plummer model with a half-mass radius of $r_{0.5} = 0.1$ pc. The stellar masses are denoted by different symbols as defined in the key. Mass segregation develops within the energy-equipartition time scale $t_{\text{eq}} < 0.4$ Myr (Kroupa 2008) such that the low-mass stars expand outward. The snapshot shows the system at a time of 1.88 Myr. By this time, a massive star system has been expelled from the cluster by a relatively gentle (about 2 km/s) cluster-potential–star recoil. The expelled massive star, which is an equal-mass binary formed after an exchange encounter within the cluster core, has a very wide intermediate-mass companion as well as a very wide M dwarf companion. It is easy to misinterpret such a hierarchical multiple system, located more than 4 pc away from a compact young cluster, to be an ideal candidate for isolated massive star formation

have, however, shown that a substantial number of even the best candidates for such an isolated population have high space motions (Ramspeck et al. 2001) which are best understood as the result of energetic stellar-dynamical ejections when massive binary systems interact in the cores of star clusters in star-forming regions. This interpretation poses important constraints on the initial properties of OB binary systems (Clarke and Pringle 1992; Kroupa 2001a; Pflamm-Altenburg and Kroupa 2006; Gvaramadze and Gualandris 2011). Still, de Wit et al. (2004, 2005) found that $4\% \pm 2\%$ of all O stars may have formed outside any cluster environment.



■ Fig. 4-7

Estimates of the percentage of massive stars formed in isolation as a function of publication year. The present-day (2011) estimate of the fraction of O stars which cannot be traced back to their birth clusters is at the lower limit of what is expected from the two-step mechanism (*shaded area*) (From Gvaramadze et al. 2012)

This percentage, however, had to be reduced twice (► Fig. 4-7) because Schilbach and Röser (2008) showed that 6 out of 11 stars that apparently formed in isolation can be back-traced to their parent clusters. Moreover, Gvaramadze and Bomans (2008) demonstrated that one of the best examples for isolated Galactic high-mass star formation (de Wit et al. 2005), the star HD 165319, has a bow shock and is thus a runaway star, most likely ejected from the young massive star cluster NGC 6611. This further reduces the percentage of massive stars possibly formed in isolation, bringing it to $1.5\% \pm 0.5\%$. And finally, using the WISE data, Gvaramadze et al. (2012) discovered a bow shock generated by one more star (HD 48279) from the sample of the best examples for isolated Galactic high-mass star formation. Correspondingly, the percentage of massive stars which may have formed in isolation is reduced to $1.0\% \pm 0.5\%$. Another example of a possible very massive star having formed in isolation is VFTS 682 which is an about $150 M_{\odot}$ heavy star located about 30 pc in projection from the star burst cluster R136 in the Large Magellanic Cloud (Bestenlehner et al. 2011). However, realistic binary-rich N -body models of initially mass-segregated R136-type clusters show that such massive stars are ejected with the observed velocities in all computations therewith readily allowing VFTS 682 to be interpreted as a slow runaway from R136 (Banerjee et al. 2012).

Massive stars may perfectly appear to have formed in isolation despite originating in clusters: The “two-step-ejection scenario,” which places massive stars outside their parent cluster such that they may fake isolated formation, has been presented by Pflamm-Altenburg and Kroupa (2010). This is based on massive binaries first being dynamically ejected from their parent star cluster. The subsequent type II supernova explosion then places the remaining massive star on a random trajectory such that it can nearly never be traced back to its parent star cluster. This is necessarily the case for 1–4% of all O stars assuming all massive stars to form in star clusters which obey the $m_{\max} - M_{\text{ecl}}$ relation. Further, if the ejected O star system consists of a tight inner binary with an outer companion the Kozai mechanism is likely to force the

inner binary into coalescence leading to a rejuvenated massive star (a massive blue straggler). When the outer companion explodes the massive blue straggler would be released in a random direction such that neither its age nor its motion would allow it to be traced to its cluster of origin.

Thus, as shown above, after excluding all observed O stars that can be traced back to young star clusters as well as those with bow shocks, the current observational evidence for the possible existence of isolated O star formation amounts to 1% of all known O stars. This is at the lower limit of the expected two-step ejection fraction of O stars that cannot be traced back to their cluster of origin. There is therefore no meaningful evidence for the formation of massive stars in isolation. The hypothesis that the isolated formation of massive stars can be a significant contributor to the population of massive stars such that the IMF of a whole galaxy effectively becomes a purely probabilistic invariant distribution function can formally be negated by noting that the observed IMF is too invariant (☛ Fig. 4-27). That is, an isolated massive star would correspond to an IMF that significantly differs, by chance, from the theoretical parent distribution function. But for each such extreme case there would be many more cases of coeval stellar populations that by chance have “strange” IMFs. There is no observational evidence whatsoever for this: all known resolved stellar populations are canonical.

Main Result

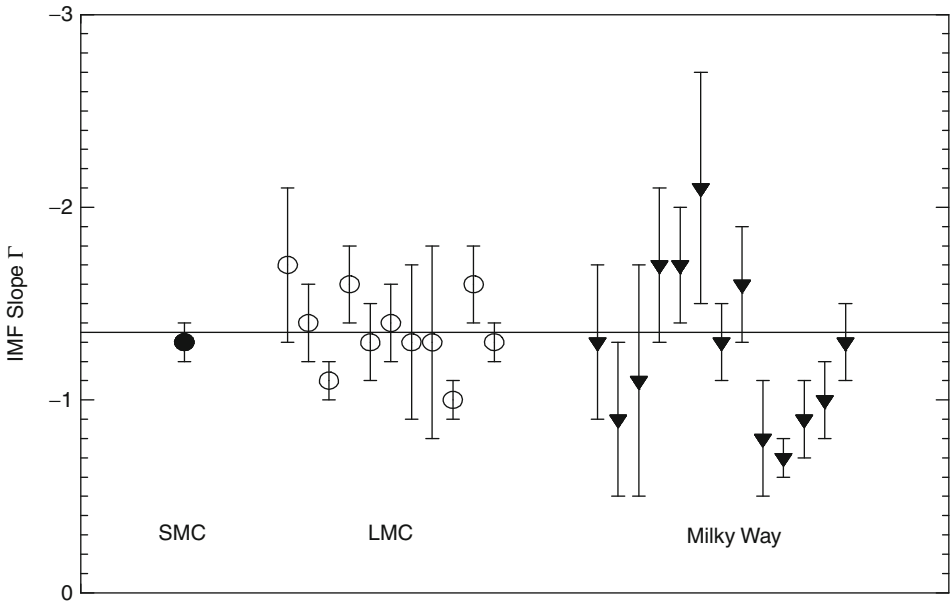
The observed field massive stars are the necessary outcome of the formation of massive stars as multiple systems in the cores of embedded clusters. There is no convincing evidence for the formation of isolated massive stars.

5 The IMF of Massive Stars

In what follows, the IMF power law indices are α_1 for $0.07 \leq m/M_\odot < 0.5$, α_2 for $0.5 \leq m/M_\odot \lesssim 1$, α_3 for $1 \lesssim m/M_\odot$.

Studying the distribution of massive stars ($\gtrsim 10 M_\odot$) is complicated because they radiate most of their energy at far-UV wavelengths that are not accessible from the Earth and through their short main-sequence lifetimes, τ , that remove them from star-count surveys. For example, a $85 M_\odot$ star cannot be distinguished from a $40 M_\odot$ star on the basis of M_V alone (Massey 2003). Constructing $\Psi(M_V)$ in order to arrive at $\Xi(m)$ for a mixed-age population does not lead to success if optical or even UV bands are used. Instead, spectral classification and broadband photometry for estimation of the reddening on a star-by-star basis has to be performed to measure the effective temperature, T_{eff} , and the bolometric magnitude, M_{bol} , from which m is obtained allowing the construction of $\Xi(m)$ directly (whereby $\Psi(M_{\text{bol}})$ and $\Xi(m)$ are related by (☛ 4.1)). Having obtained $\Xi(m)$ for a population under study, the IMF follows by applying (☛ 4.3) after evolving each measured stellar mass to its initial value using theoretical stellar evolution tracks.

Massey (2003) stresses that studies that only rely on broadband optical photometry consistently arrive at IMFs that are significantly steeper with $\alpha_3 \approx 3$, rather than $\alpha_3 = 2.2 \pm 0.1$ found for a wide range of stellar populations using spectroscopic classification. Indeed, the application of the same methodology by Massey on a number of young populations of different metallicity and density shows a remarkable uniformity of the IMF above about $10 M_\odot$ (☛ Fig. 4-8).



■ Fig. 4-8

The IMF slope $\Gamma = 1 - \alpha_3$ determined in a homogeneous manner for OB associations and clusters in the MW, LMC, and SMC. The Small Magellanic Cloud (SMC) has a metallicity $Z = 0.005$ ($[\text{Fe}/\text{H}] \approx -0.6$), the Large Magellanic Cloud (LMC) has $Z = 0.008$ ($[\text{Fe}/\text{H}] \approx -0.4$), and the Milky Way (MW) has $Z = 0.018$ ($[\text{Fe}/\text{H}] \approx -0.05$) within a distance of 3 kpc around the Sun (with kind permission from Massey (2003))

The available IMF measurements do not take into account the bias through unresolved systems which may, in principle, be substantial since the proportion of multiple stars is higher for massive stars than for low-mass Galactic-field stars (e.g., Duchêne et al. 2001). For example, in the Orion Nebula Cluster (ONC), each massive star has, on average, 1.5 companions (Preibisch et al. 1999), while in the cluster NGC 6231, García and Mermilliod (2001) find that 80% of all O stars are radial-velocity binaries.

However, Maíz Apellániz (2008) and Weidner et al. (2009) demonstrate that the observed IMF for massive stars is not affected significantly by unresolved multiple systems: The models, where initial masses are derived from the luminosity and color of unresolved multiple systems, show that even under extreme circumstances (100% binaries or higher order multiples), the difference between the power-law index of the mass function (MF) of all stars and the observed MF is small ($\Delta\alpha \lesssim 0.1$). Thus, if the observed IMF has the Salpeter index $\alpha_3 = 2.35$, then the true stellar IMF has an index not flatter than $\alpha_3 = 2.25$.

Massive main-sequence stars have substantial winds flowing outward with velocities of a few 100 to a few 1,000 km/s (Kudritzki and Puls 2000). For example, $10^{-6.5} < \dot{m} < 10^{-6} M_\odot/\text{year}$ for $m = 35 M_\odot$ with main-sequence lifetime $\tau = 4.5$ Myr (García-Segura et al. 1996a) and $10^{-5.6} < \dot{m} < 10^{-5.8} M_\odot/\text{year}$ for $m = 60 M_\odot$ with $\tau = 3.45$ Myr (García-Segura et al. 1996b). More problematical is that stars form rapidly rotating and are sub-luminous as a result of reduced internal pressure. But they decelerate significantly during their main-sequence lifetime owing to the angular-momentum loss through their winds and become more luminous more

rapidly than nonrotating stars (Maeder and Meynet 2000). For ages less than 2.5 Myr, the models deviate only by 5–13% from each other in mass, luminosity, or temperature (Weidner and Kroupa 2006). Large deviations are evident for advanced stages of evolution though because of the sensitivity to the different treatment of the stellar physics.

The mass–luminosity relation for a population of stars that have a range of ages is broadened, making mass estimates from M_{bol} uncertain by up to 50% (Penny et al. 2001), a bias that probably needs to be taken into account more thoroughly in the derivations of the IMF. Another problem is that $m \gtrsim 40 M_{\odot}$ stars may finish their assembly after burning a significant proportion of their central H so that a zero-age main sequence may not exist for massive stars (Maeder and Behrend 2002). However, the agreement between slowly rotating tidally locked massive O-type binaries with standard nonrotating theoretical stellar models is very good (Penny et al. 2001).

Main Results

The IMF of massive stars is well described by a *Salpeter/Massey* slope, $\alpha_3 = 2.3$, independent of environment as deduced from resolved stellar populations in the Local Group of galaxies. Unresolved multiple stars do not significantly affect the massive-star power-law index.

A note to the statement that $\alpha_3 = 2.3$ is independent of environment: This is strictly only valid for star formation with densities less than about $10^5 M_{\odot}/\text{pc}^3$ and for metallicities of $[\text{Fe}/\text{H}] \gtrsim -2$ as are observed in the Local Group of galaxies. Evidence has emerged that star formation at higher densities leads to top-heavy IMFs (see [Fig. 4-31](#) below).

6 The IMF of Intermediate-Mass Stars

Intermediate-mass ($\approx 1\text{--}8 M_{\odot}$) stars constitute a complicated sample to deal with in terms of inferring their mass function from star counts in the Galactic field as well as in star clusters. Their lifetimes are comparable to the age of the Galactic disk down to the lifetime of typical open clusters (a few 10^8 year, [Fig. 4-26](#) below). Also, the distribution function of multiple systems changes in this mass range ([Sect. 2.6](#)). Corrections of their luminosities for stellar evolution and binarity are thus more challenging than in the other mass ranges. Also, the diffusion of stellar orbits within the MW disk away from the birth orbits has a comparable time scale (the dynamical period of the Milky Way) such that an ensemble of intermediate-mass stars does not have an as well constrained Galactic-disk thickness as the massive (≈ 50 pc) or low-mass (≈ 500 pc) stars. This affects the combination of the essentially two-dimensional (in the Galactic disk plane) star counts of massive stars with the three-dimensional (solar neighborhood within a few hundred pc) star counts of late-type main sequence stars to a common density of stars in dependence of stellar luminosity. This issue is dealt with excellently by Scalo (1986). The resulting constraints on the IMF in this mass range are rather uncertain. Indeed, in [Fig. 4-26](#), it is evident that the scatter of deduced α -indices is very large in this mass range, and the analysis by Scalo (1986) may even suggest a discontinuity in the Galactic-disk IMF in this mass range.

The stellar IMF has $\alpha_2 \approx 2.3$ for stars with $0.5 < m/M_{\odot} < 1$ (main results on p. 177) and $\alpha_3 \approx 2.3$ for $m \gtrsim 8 M_{\odot}$ ([Sect. 5](#)) such that $\alpha_3 \approx 2.3$ for $1 \lesssim m/M_{\odot} \lesssim 8$ appears natural. However,

as noted in [Sect. 2.6](#), the initial binary properties appear to be different for $m < \text{few } M_{\odot}$ in comparison to those for $m > \text{few } M_{\odot}$ implying open question I on p. 143.

Here, it is assumed that the IMF is continuous across this mass range, and thus attention is given to the massive ([Sect. 5](#)) and low-mass stars ([Sect. 7](#)). But in view of the discontinuity issue uncovered on p. 178 for stars and brown dwarfs, the continuity assumption made here needs to be kept in mind.

7 The IMF of Low-Mass Stars (LMSs)

Here, stars with $0.1 \lesssim m/M_{\odot} \lesssim 1$ (the LMSs) are discussed. They are late-type main sequence stars which constitute the vast majority of all stars in any known stellar population ([Table 4-1](#) below). Also, their initial binary properties follow rather simple rules ([Sect. 2.6](#)). Very low-mass stars (VLMSs) with $m \lesssim 0.15 M_{\odot}$ are the subject of [Sect. 8](#).

There are three well-tried approaches to determine $\Psi(M_V)$ in ([4.1](#)). The first two are applied to Galactic-field stars (parallax- and flux-limited star-counts) and the third to star clusters (establishment of members). The sample of Galactic-field stars close to the Sun (typically within 5–20 pc distance depending on m) is especially important because it is the most complete and well-studied stellar sample at our disposal.

7.1 Galactic-Field Stars and the Stellar Luminosity Function

Galactic-field stars have an average age of about 5 Gyr and represent a mixture of many star-formation events. The IMF deduced for these is therefore a time-cumulated composite IMF (i.e., the IGIMF, [Sect. 13.1](#)). For $m \lesssim 1.3 M_{\odot}$, the composite IMF equals the stellar IMF according to the presently available analysis, and so it is an interesting quantity for at least two reasons: For the mass budget of the Milky Way disk and as a benchmark against which the IMFs measured in present- and past-occurring star-formation events can be compared to distill possible variations about the mean.

The *first and most straightforward method* consists of creating a local volume-limited catalogue of stars yielding the nearby LF, $\Psi_{\text{near}}(M_V)$. Completeness of the modern *Jahreiss–Gliese Catalogue of Nearby Stars* extends to about 25 pc for $m \gtrsim 0.6 M_{\odot}$, trigonometric distances having been measured using the Hipparcos satellite, and only to about 5 pc for less massive stars for which we still rely on ground-based trigonometric parallax measurements.¹³ The advantage of the LF, $\Psi_{\text{near}}(M_V)$, created using this catalogue, is that virtually all companion stars are known, that it is truly distance limited, and that the distance measurements are direct.

¹³Owing to the poor statistical definition of $\Psi_{\text{near}}(M_V)$ for $m \lesssim 0.5 M_{\odot}$, $M_V \gtrsim 10$, it is important to increase the sample of nearby stars, but controversy exists as to the maximum distance to which the LMS census is complete. Using spectroscopic parallax, it has been suggested that the local census of LMSs is complete to within about 15% to distances of 8 pc and beyond (Reid and Gizis 1997). However, Malmquist bias allows stars and unresolved binaries to enter such a flux-limited sample from much larger distances (Kroupa 2001c). The increase of the number of stars with distance using trigonometric distance measurements shows that the nearby sample becomes incomplete for distances larger than 5 pc and for $M_V > 12$ (Jahreiss 1994; Henry et al. 1997). The incompleteness in the northern stellar census beyond 5 pc and within 10 pc amounts to about 35% (Jao et al. 2003), and further discovered companions (e.g., Delfosse et al. 1999; Beuzit et al. 2004) to known primaries in the distance range $5 < d < 12$ pc indeed suggest that the extended sample may not yet be complete. Based on the work by Reid et al. (2003a, b), Luhman (2004), however, argues that the incompleteness is only about 15%.



■ Table 4-1

See ▶ Sect. 10.1. The number fraction $\eta_N = 100 \int_{m_1}^{m_2} \xi(m) dm / \int_{m_1}^{m_u} \xi(m) dm$; the mass fraction $\eta_M = 100 \int_{m_1}^{m_2} m \xi(m) dm / M_{cl}$, $M_{cl} = \int_{m_1}^{m_u} m \xi(m) dm$, in percent of main-sequence stars in the mass interval m_1 to m_2 ; and the stellar contribution, ρ^{st} , to the Oort limit and to the Galactic-disk surface mass density, $\Sigma^{st} = 2h\rho^{st}$, near to the Sun, taking $m_l = 0.07 M_\odot$, $m_u = 120 M_\odot$, and the Galactic-disk scale height $h = 250$ pc ($m < 1 M_\odot$ Kroupa et al. 1993) and $h = 90$ pc ($m > 1 M_\odot$, Scalo 1986). Results are shown for the canonical IMF (▶ 4.55) for the high-mass-star Galactic-field IMF ($\alpha_3 = 2.7$, $m > 1 M_\odot$) and for the PDMF ($\alpha_3 = 4.5$, Scalo 1986; Kroupa et al. 1993) which describes the distribution of stellar masses now populating the Galactic disk. For gas, $\Sigma^{gas} = 13 \pm 3 M_\odot/\text{pc}^2$ and remnants $\Sigma^{rem} \approx 3 M_\odot/\text{pc}^2$ (Weidemann 1990). The average stellar mass is $\bar{m} = \int_{m_1}^{m_u} m \xi(m) dm / \int_{m_1}^{m_u} \xi(m) dm$. m_{max} and M_{ecl} are calculated using equations ▶ 4.10 and ▶ 4.11. with $m_{max*} = 120M_\odot$. N_{cl} is the number of stars that have to form in a star cluster such that the most massive star in the population has the mass m_{max} . The mass of this population is M_{cl} , and the condition is $\int_{m_{max}}^{120M_\odot} \xi(m) dm = 1$ with $\int_{0.07}^{m_{max}} \xi(m) dm = N_{cl} - 1$. $\Delta M_{cl}/M_{cl}$ is the fraction of mass lost from the cluster due to stellar evolution, assuming that for $m \geq 8 M_\odot$, all neutron stars and black holes are kicked out due to an asymmetrical supernova explosion but that white dwarfs are retained (Weidemann et al. 1992). The masses of the white dwarfs are estimated as $m_{WD} = 0.109 m_{ini} + 0.394 [M_\odot]$, which is a linear fit to the masses of observed white dwarfs (Kalirai et al. 2008). The evolution time for a star of mass m_{to} to reach the turnoff age is available in ▶ Fig. 4-26. Note that brown dwarfs are not considered for any of the numbers listed in this table


Mass range (M_\odot)	η_N (%)			η_M (%)			ρ^{st}	Σ^{st}
	α_3	α_3	α_3	α_3	α_3	α_3	(M_\odot/pc^2)	(M_\odot/pc^2)
	2.3	2.7	4.5	2.3	2.7	4.5	4.5	4.5
0.07–0.5	77.71	79.39	82.38	28.42	37.96	52.90	2.17×10^{-2}	9.73
0.5–1	13.25	13.54	14.05	16.66	22.24	31.00	1.27×10^{-2}	5.70
1–8	8.45	6.87	3.57	33.44	31.62	16.01	6.56×10^{-3}	2.95
8–120	0.59	0.20	0.00	21.48	8.18	0.09	3.69×10^{-5}	1.66×10^{-2}
$\bar{m}/M_\odot =$	0.545	0.417	0.310				$\rho_{tot}^{st} = 0.041$	$\Sigma_{tot}^{st} = 18.4$
m_{max} (M_\odot)	$\alpha_3 = 2.3$		$\alpha_3 = 2.7$		m_{to} (M_\odot)	$\Delta M_{cl}/M_{cl}$ (%)		
	N_{cl}	M_{cl} (M_\odot)	N_{cl}	M_{cl} (M_\odot)		$\alpha_3 = 2.3$	$\alpha_3 = 2.7$	
1	13	3.2	15	3.8	80	2.2	0.5	
8	173	82	489	187	60	4.0	0.9	
20	601	307	2415	970	40	6.7	1.7	
40	1756	893	8839	3623	20	12.2	3.6	
60	3803	1993	21509	8885	8	21.5	8.2	
80	8023	4275	48750	20224	3	32.1	15.8	
100	20820	11236	133129	55385	1	45.2	29.9	
119	509319	277416	3.38×10^6	1.41×10^6	0.7	48.6	34.4	



The *second method* is to make deep pencil-beam surveys using photographic plates or CCD cameras to extract a few hundred low-mass stars from a hundred-thousand stellar and galactic images. This approach, pioneered by Gerry Gilmore and Neill Reid, leads to larger stellar

samples especially so since many lines of sight into the Galactic field ranging to distances of a few hundred pc to a few kpc are possible. The disadvantage of the LF, $\Psi_{\text{phot}}(M_V)$, created using this technique is that the distance measurements are indirect relying on photometric parallax. Such surveys are flux-limited rather than volume-limited, and pencil-beam surveys which do not pass through virtually the entire stellar disk are prone to Malmquist bias (Stobie et al. 1989). This bias results from a spread of luminosities of stars that have the same color because of their dispersion of metallicities and ages, leading to intrinsically more luminous stars entering the flux-limited sample and thus biasing the inferred absolute luminosities and the inferred stellar space densities. Furthermore, binary systems are not resolved in the deep surveys, or if formally resolved, the secondary is likely to be below the flux limit.



The local, *nearby LF* and the Malmquist-corrected deep *photometric LF* are displayed in  Fig. 4-9. They differ significantly for stars fainter than $M_V \approx 11.5$ which caused some controversy in the past.¹⁴ That the local sample has a spurious but significant overabundance of low-mass stars can be ruled out by virtue of the large velocity dispersion in the disk, ≈ 30 pc/Myr. Any significant overabundance of stars within a sphere with a radius of 30 pc would disappear within 1 Myr and cannot be created by any physically plausible mechanism from a population of stars with stellar ages spanning the age of the Galactic disk. The shape of $\Psi_{\text{phot}}(M_V)$ for $M_V \gtrsim 12$ is confirmed by many independent photometric surveys. That all of these could be making similar mistakes, such as in color transformations, becomes unlikely on consideration of the LFs constructed for completely independent stellar samples, namely, star clusters ( Fig. 4-10).

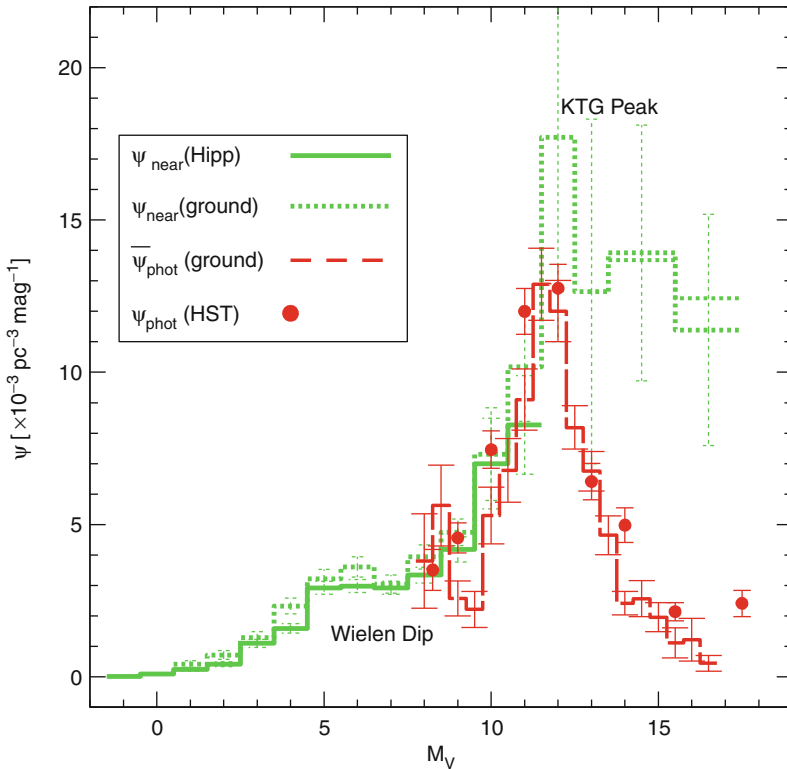
7.2 The Stellar Mass–Luminosity Relation

The MF is related to the LF via the derivative of the stellar mass–luminosity relation ( 4.1).

 Equation 4.1 shows that any nonlinear structure in the MLR is mapped into observable structure in the LF, provided the MF does not have a compensating structure. Such a conspiracy is implausible because the MF is defined through the star-formation process, but the MLR is a result of the internal constitution of stars. The MLR and its derivative are shown in  Fig. 4-11. It is apparent that the slope is very small at faint luminosities leading to large uncertainties in the MF near the hydrogen burning mass limit.

The physics underlying the nonlinearities of the MLR are due to an interplay of changing opacities, the internal stellar structure, and the equation of state of the matter deep inside the

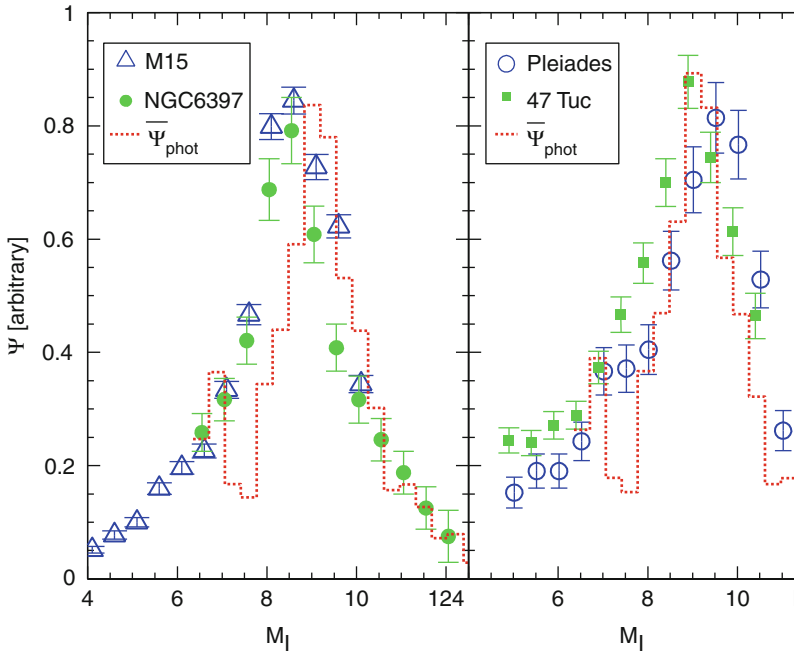
¹⁴This controversy achieved a maximum in 1995, as documented in Kroupa (1995a). The discrepancy evident in  Fig. 4-9 between the nearby LF, Ψ_{near} , and the photometric LF, Ψ_{phot} , invoked a significant dispute as to the nature of this discrepancy. On the one hand (Kroupa 1995a), the difference is thought to be due to unseen companions in the deep but low-resolution surveys used to construct Ψ_{phot} , with the possibility that photometric calibration for VLMSs may remain problematical so that the exact shape of Ψ_{phot} for $M_V \gtrsim 14$ is probably uncertain. On the other hand (Reid and Gizis 1997), the difference is thought to come from nonlinearities in the $V - I, M_V$ color–magnitude relation used for photometric parallax. Taking into account such structure, it can be shown that the photometric surveys underestimate stellar space densities so that Ψ_{phot} moves closer to the extended estimate of Ψ_{near} using a sample of stars within 8 pc or further. While this is an important point, the extended Ψ_{near} is incomplete (see footnote 13 on p. 160) and theoretical color–magnitude relations do not have the required degree of nonlinearity (e.g., Fig. 7 in Bochanski et al. 2010). The observational color–magnitude data also do not conclusively suggest a feature with the required strength (Baraffe et al. 1998). Furthermore, Ψ_{phot} agrees almost perfectly with the LFs measured for star clusters of solar and population II metallicity for which the color–magnitude relation is not required ( Fig. 4-10) so that nonlinearities in the color–magnitude relation cannot be the dominant source of the discrepancy



■ Fig. 4-9

Stellar luminosity functions (LFs) for solar-neighborhood stars. The photometric LF corrected for Malmquist bias and at the midplane of the Galactic disk (Ψ_{phot}) is compared with the nearby LF (Ψ_{near}). The average, ground-based $\bar{\Psi}_{\text{phot}}$ (dashed histogram, data predating 1995, is confirmed (Kroupa 1995a) by Hubble Space Telescope (HST) star-count data which pass through the entire Galactic disk and are thus less prone to Malmquist bias (solid dots). The ground-based volume-limited trigonometric-parallax sample (dotted histogram) systematically overestimates Ψ_{near} due to the Lutz–Kelker bias, thus lying above the improved estimate provided by the Hipparcos-satellite data (solid histogram, Jahreiß and Wielen 1997; Kroupa 2001c). The Lutz–Kelker bias (Lutz and Kelker 1973) arises in trigonometric-parallax-limited surveys because the uncertainties in parallax measurements combined with the nonlinear increase of the number of stars with reducing parallax (increasing distance) lead to a bias in the deduced number density of stars when using trigonometric-parallax-limited surveys. The depression/plateau near $M_V = 7$ is the *Wielen dip*. The maximum near $M_V \approx 12$, $M_I \approx 9$ is the *KTG peak*. The thin dotted histogram at the faint end indicates the level of refinement provided by additional stellar additions (Kroupa 2001c), demonstrating that even the immediate neighborhood within 5.2 pc of the Sun probably remains incomplete at the faintest stellar luminosities. Which LF is the relevant one for constraining the MF? Kroupa et al. (1993) uniquely used both LFs simultaneously to enhance the constraints. See text

stars. Starting at high masses ($m \gtrsim \text{few } M_{\odot}$), as the mass of a star is reduced, H^{-} opacity becomes increasingly important through the short-lived capture of electrons by H-atoms, resulting in reduced stellar luminosities for intermediate and low-mass stars. The $m(M_V)$ relation becomes less steep in the broad interval $3 < M_V < 8$ leading to the *Wielen dip* (☛ Fig. 4-9). The $m(M_V)$

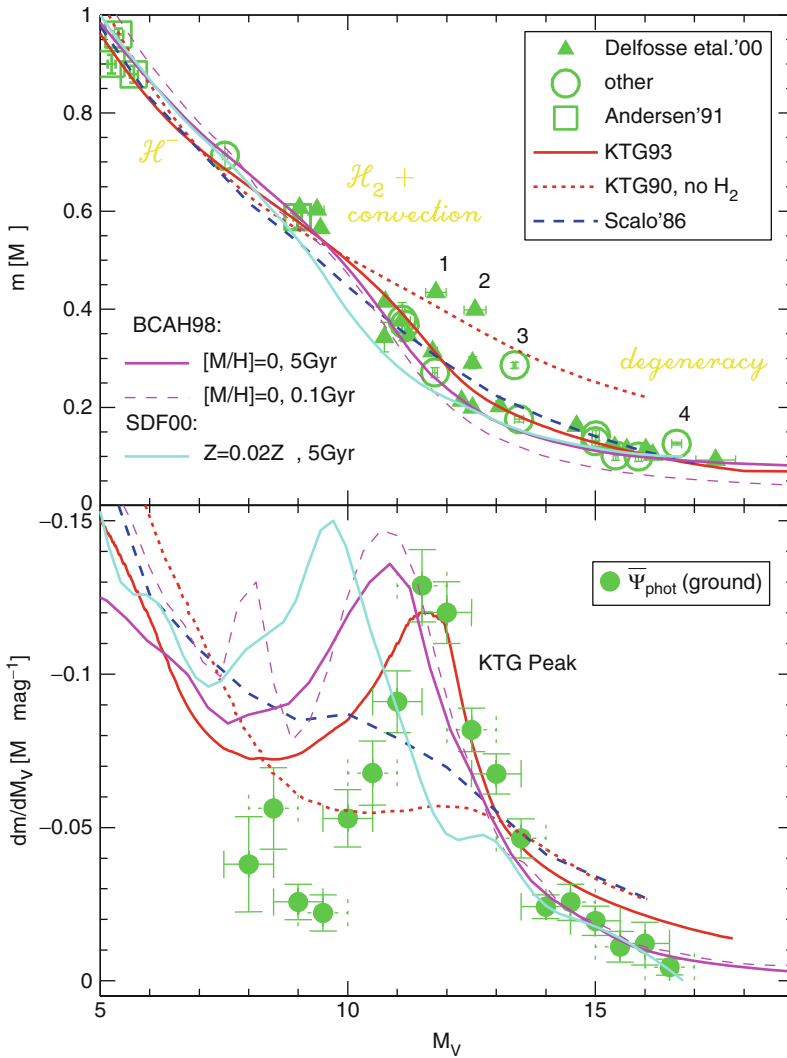


■ Fig. 4-10

I-band LFs of stellar systems in four star clusters: globular cluster (GC) *M15* (de Marchi and Paresce 1995a, distance modulus $\Delta m = m - M = 15.25$ mag), GC *NGC 6397* (Paresce et al. 1995, $\Delta m = 12.2$), young Galactic cluster *Pleiades* (Hambly et al. 1991, $\Delta m = 5.48$), and GC *47 Tuc* (de Marchi and Paresce 1995b, $\Delta m = 13.35$). The dotted histogram is $\bar{\Psi}_{\text{phot}}(M_I)$ from Fig. 4-9, transformed to the *I*-band using the linear color–magnitude relation $M_V = 2.9 + 3.4(V - I)$ (Kroupa et al. 1993), and $\Psi_{\text{phot}}(M_I) = (dM_V/dM_I) \Psi_{\text{phot}}(M_V)$ (Eq. 4.2). The KTG peak is very pronounced in all LFs. It is due to an extremum in the derivative of the MLR (Fig. 4-11)

relation steepens near $M_V = 10$ because the formation of H_2 in the very outermost layer of main-sequence stars causes the onset of convection up to and above the photo-sphere leading to a flattening of the temperature gradient and therefore to a larger effective temperature as opposed to an artificial case without convection but the same central temperature. This leads to brighter luminosities and full convection for $m \lesssim 0.35 M_\odot$. The modern Delfosse data beautifully confirm the steepening in the interval $10 < M_V < 13$ predicted in 1990. In Fig. 4-11, the dotted MLR demonstrates the effects of suppressing the formation of the H_2 molecule by lowering its dissociation energy from 4.48 to 1 eV (Kroupa et al. 1990, hereinafter KTG). The $m(M_V)$ relation flattens again for $M_V > 14$, $m < 0.2 M_\odot$ as degeneracy in the stellar core becomes increasingly important for smaller masses limiting further contraction (Hayashi and Nakano 1963; Chabrier and Baraffe 1997). Therefore, owing to the changing conditions within the stars with changing mass, a pronounced local maximum in $-dm/dM_V$ results at $M_V \approx 11.5$, postulated by KTG to be the origin of the maximum in Ψ_{phot} near $M_V = 12$.

The implication that the LFs of all stellar populations should show such a feature, although realistic metallicity-dependent stellar models were not available yet, was noted (Kroupa et al. 1993). The subsequent finding that all known stellar populations have the KTG peak



■ Fig. 4-11

The mass–luminosity relation (MLR, *upper panel*) and its derivative (*lower panel*) for late-type stars. *Upper panel:* The observational data (solid triangles and open circles, Delfosse et al. 2000; open squares, Andersen 1991) are compared with the empirical MLR of Scalo (1986) and the semiempirical KTG93-MLR tabulated in Kroupa et al. (1993). The under-luminous data points 1,2 are GJ2069Aa,b and 3,4 are Gl791.2A,B. All are probably metal-rich by ≈ 0.5 dex (Delfosse et al. 2000). Theoretical MLRs from Baraffe et al. (1998) (BCAH98) and Siess et al. (2000) (SDF00) are also shown. The observational data (Andersen 1991) show that $\log_{10}[m(M_\odot)]$ is approximately linear for $m > 2 M_\odot$. See also Malkov et al. (1997). *Lower panel:* The derivatives of the same relations plotted in the *upper panel* are compared with $\bar{\Psi}_{\text{phot}}$ from Fig. 4-9. Note the good agreement between the location, amplitude, and width of the KTG peak in the LF and the extremum in dm/dM_V

(► *Figs. 4-9* and ► *4-10*) constitutes one of the *most impressive achievements of stellar-structure theory*. Different theoretical $m(M_V)$ relations have the maximum in $-dm/dM_V$ at different M_V , suggesting the possibility of testing stellar structure theory near the critical mass $m \approx 0.35 M_\odot$, where stars become fully convective (Kroupa and Tout 1997; Brocato et al. 1998). But since the MF also defines the LF, the shape and location cannot be unambiguously used for this purpose unless it is postulated that the IMF is invariant. Another approach to test stellar models is by studying the deviations of observed $m(M_V)$ data from the theoretical relations (► *Fig. 4-12*).

A study of the position of the maximum in the *I*-band LF has been undertaken by von Hippel et al. (1996) and Kroupa and Tout (1997) finding that the observed position of the maximum shifts to brighter magnitude with decreasing metallicity, as expected from theory (► *Figs. 4-13* and ► *4-14*).

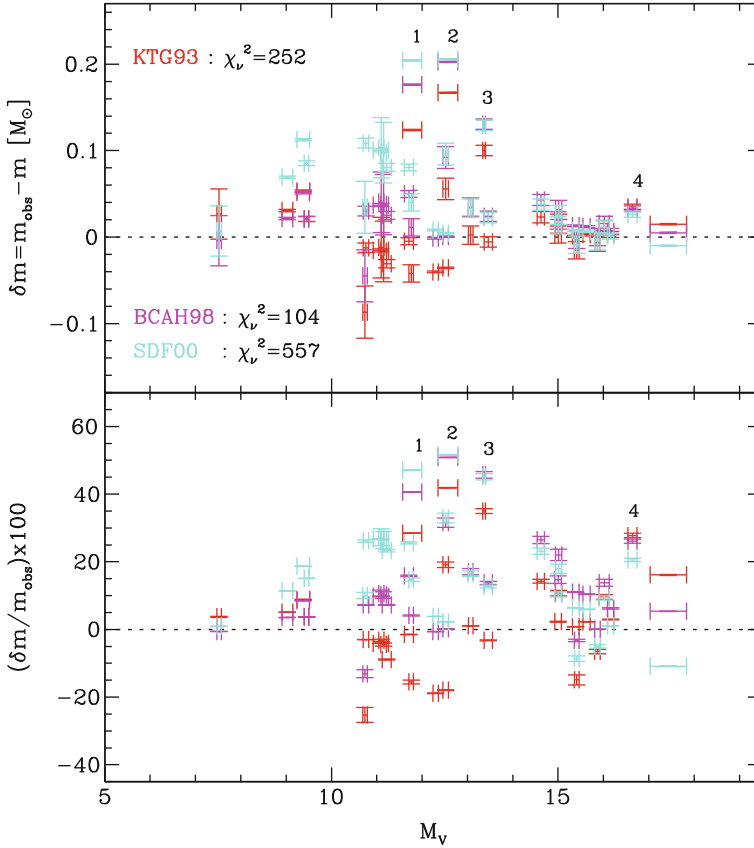
7.3 Unresolved Binary Stars and the Solar-Neighborhood IMF

In addition to the nonlinearities in the $m(M_p)$ relation, unresolved multiple systems affect the MF derived from the photometric LF, in particular since no stellar population is known to exist that has a binary proportion smaller than about 50%, apart possibly from dynamically highly evolved globular clusters and open clusters (Sollima et al. 2007, 2010; Marks et al. 2011).

Suppose an observer sees 100 systems. Of these 40, 15 and 5 are binary, triple, and quadruple, respectively, these being realistic proportions. There are thus 85 companion stars which the observer is not aware of if none of the multiple systems are resolved. Since the distribution of secondary masses is not uniform but typically increases with decreasing mass for F-, G- and K-type primaries (it decreases for M-type primaries, Malkov and Zinnecker 2001; Marks et al. 2011), the bias is such that low-mass stars are significantly underrepresented in any survey that does not detect companions (Kroupa et al. 1991; Malkov and Zinnecker 2001). Also, if the companion(s) are bright enough to affect the system luminosity noticeably then the estimated photometric distance will be too small, artificially enhancing inferred space densities which are, however, mostly compensated for by the larger distances sampled by binary systems in a flux-limited survey, together with an exponential density falloff perpendicular to the Galactic disk (Kroupa 2001c). A faint companion will also be missed if the system is formally resolved, but the companion lies below the flux limit of the survey.

Comprehensive star-count modeling of the solar neighborhood that incorporates unresolved binary systems, metallicity, and age spreads and the density falloff perpendicular to the Galactic disk with appropriate treatment of Malmquist and Lutz–Kelker bias show that the IMF, from which the solar neighborhood populations within a few pc and a few hundred pc stem, can be *unified with one MF* which is a two-part power law with $\alpha_1 = 1.3 \pm 0.5$, $0.07 < m/M_\odot \leq 0.5$, $\alpha_2 \approx 2.2$, $0.5 < m/M_\odot \leq 1$, a result obtained for two different MLRs (Kroupa et al. 1993; Kroupa 2001c). The index α_2 is constrained tightly owing to the well-constrained Ψ_{near} , the well-constrained empirical MLR in this mass range and because unresolved binary systems do not significantly affect the solar-neighborhood LF in this mass range because primaries with $m \gtrsim 1 M_\odot$ are rare and are not sampled. The stellar sample in the mass range $0.5\text{--}1 M_\odot$ is therefore complete.

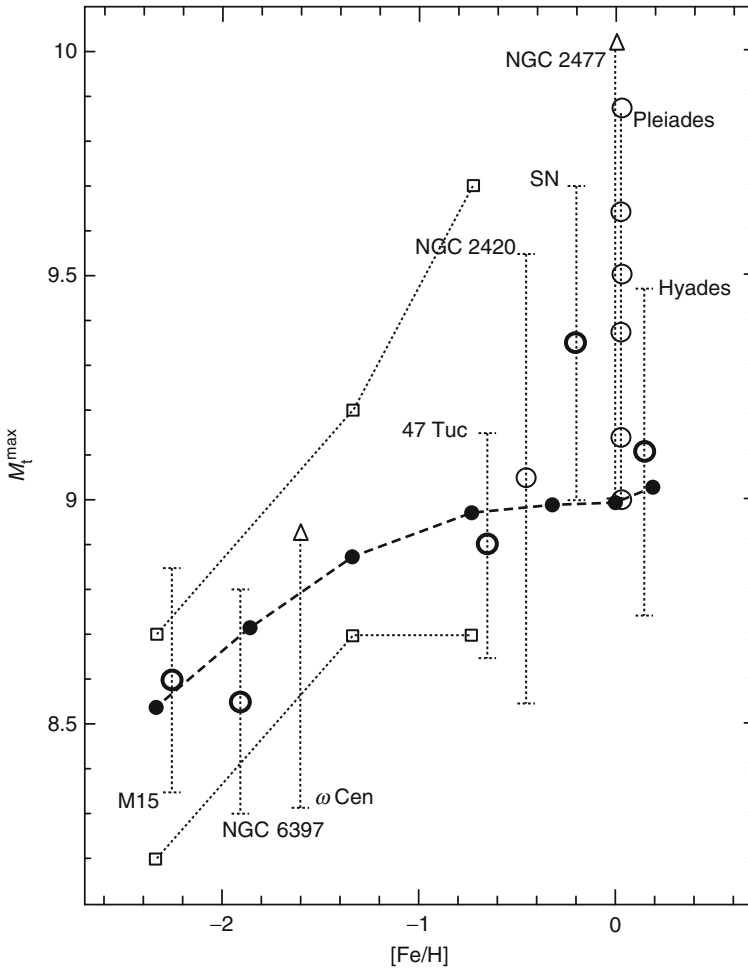
► *Figure 4-15* demonstrates models of the individual-star and system LFs for the KTG93 MLR shown in ► *Fig. 4-11*. The significant difference between the individual-star and system LFs is evident, being most of the explanation of the disputed discrepancy between the observed



■ Fig. 4-12

Deviations of the MLRs ($\delta m = m_{\text{obs}} - m(M_V)$) from the empirical data, m_{obs} , with errors ϵ_m shown in [Fig. 4-11](#) in M_{\odot} (*upper panel*) and in percent (*lower panel* with uncertainties $-(m(M_V)/m_{\text{obs}}^2) \times \epsilon_m$). Colors refer to the models of [Fig. 4-11](#). Reduced χ^2_{ν} ($\nu = 26$ for 31 data points, ignoring the four outliers) values indicate the formal goodness of fit. Formally, none of the MLRs available is an acceptable model for the data. This is not alarming though because the models are for a single-metallicity, single-age population while the data span a range of metallicities and ages typical for the solar neighborhood stellar population, as signified by $\delta m \gg \epsilon_m$ in most cases. The χ^2_{ν} values confirm that the BCAH98 models (Baraffe et al. 1998) and the semiempirical KTG93 MLR (Kroupa et al. 1993) provide the best-matching MLRs. Note that the KTG93 MLR was derived from mass-luminosity data prior to 1980, but by using the shape of the peak in $\Psi_{\text{phot}}(M_V)$ as an additional constraint, the constructed MLR became robust. The *lower panel* demonstrates that the deviations of observational data from the model MLRs are typically much smaller than 30%, excluding the putatively metal-rich stars (1–4)

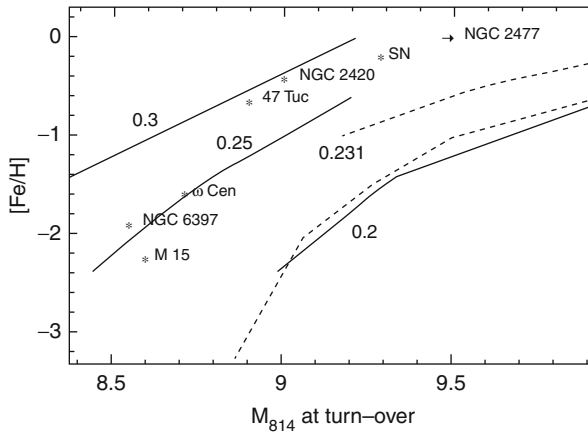
Ψ_{near} and Ψ_{phot} . Note though that the observed photometric LF contains triple and quadruple systems that are not accounted for by the model. Note also that the photometric LF has been corrected for Malmquist bias and so constitutes the system LF in which the broadening due to a metallicity and age spread and photometric errors has been largely removed. It is therefore directly comparable to the model system LF, and both indeed show a very similar



■ Fig. 4-13

The position of the maximum in $-dm/dM_i$ as a function of metallicity of the theoretical mass-luminosity data of Kroupa and Tout (1997) is shown as *solid dots*. The *open squares* represent bounds by the stellar-structure models of D'Antona and Mazzitelli (1996), and the *open circles* are observational constraints for different populations (e.g., SN for the composite solar-neighborhood population, Pleiades for the simple population of an intermediate-age cluster). *Thick circles* are more certain than the *thin circles*, and for the Pleiades, a sequence of positions of the LF-maximum is given, from *top to bottom*, with the following combinations of (distance modulus, age): (5.5, 70 Myr), (5.5, 120 Myr), (5.5, main sequence), (6, 70 Myr), (6, 120 Myr), and (6, main sequence). For more details, see Kroupa and Tout (1997)

KTG peak. The observed nearby LF, on the other hand, has not been corrected for the metallicity and age spread nor for trigonometric distance errors, and so it appears broadened. The model individual-star LF, in contrast, does not, by construction, incorporate these and thus appears with a more pronounced maximum. Such observational effects can be incorporated rather easily



■ Fig. 4-14

Similar to [Fig. 4-13](#) but from von Hippel et al. (1996), their Fig. 5: The absolute *I*-band-equivalent magnitude of the maximum in the LF as a function of metallicity for different populations. The *solid* and *dashed* lines are loci of constant mass (0.2, 0.231, 0.3 M_{\odot}) according to theoretical stellar structure calculations. See von Hippel et al. (1996) for more details

into full-scale star-count modeling (Kroupa et al. 1993). The deviation of the model system LF from the observed photometric LF for $M_V \gtrsim 14$ may indicate a change of the pairing properties of the VLMS or BD population ([Sect. 8](#)).

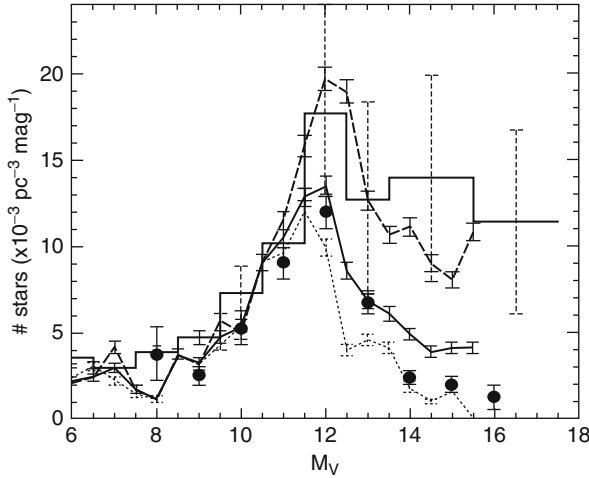
Since the nearby LF is badly defined statistically for $M_V \gtrsim 13$, the resulting model, such as shown in [Fig. 4-15](#), is a *prediction* of the true, individual-star LF that should become apparent once the immediate solar-neighborhood sample has been enlarged significantly through the planned space-based astrometric survey Gaia (Gilmore et al. 1998), followed by an intensive follow-up imaging and radial-velocity observing program scrutinizing every nearby candidate for unseen companions (Kroupa 2001c). Despite such a monumental effort, the structure in $\Psi_{\text{near}}^{\text{GAIA}}$ will be smeared out due to the metallicity and age spread of the local stellar sample, a factor to be considered in detail.

Main Results

The universal structure in the stellar LF of main sequence stars (the Wielen dip and the KTG peak) is well understood. It is due to nonlinearities in the stellar mass–luminosity relation. Binary systems have a highly significant effect on the LF of late-type stars. The solar-neighborhood IMF which unifies Ψ_{near} and Ψ_{phot} has $\alpha_1 = 1.3$ (0.07–0.5 M_{\odot}) and $\alpha_2 = 2.2$ (0.5–1 M_{\odot}).

7.4 Star Clusters

Star clusters less massive than about $M = 10^5 M_{\odot}$ and to a good degree of approximation also those with $M > 10^5 M_{\odot}$ offer populations that are co-eval, equidistant, and that have the same chemical composition. But, seemingly as a compensation of these advantages, the extraction



■ Fig. 4-15

Comparison of the model field luminosity function (curves) of a single-metallicity and single-age population that is without measurement errors, with observations in the photometric V-band (a comparison of the corresponding LFs in bolometric magnitudes can be found in Kroupa 1995e). The model assumes the standard or canonical stellar IMF, (☛ 4.55) below. The model single star luminosity function is normalized to the nearby luminosity function at $M_V \approx 10$, $M_{\text{bol}} \approx 9$, giving the normalization constant in the MF k (☛ 4.5), and the plot shows $k \Psi_{\text{mod,sing}}$ (long dashed curve), $k \Psi_{\text{mod,sys}}(t = 0)$ (dotted curve, 100% birth binary fraction in dynamically unevolved embedded clusters, see ☛ Fig. 4-16) without pre-main sequence brightening, and $k \Psi_{\text{mod,sys}}(t = 1 \text{ Gyr})$ (solid curve, 48% surviving binary fraction in dissolved clusters, see ☛ Fig. 4-16). Note that the solid curve is the luminosity function for a realistic model of the Galactic field population of systems consisting of 48% binaries (which result from disruption of the 100% binary birth population of ☛ Sect. 2.6 in the embedded clusters) which have a period distribution consistent with the observed G-, K-, and M-dwarf period distribution, the mass ratio distributions for G-dwarf systems as observed (Duquennoy and Mayor 1991), and the overall mass-ratio distribution given by Fig. 2 in Kroupa et al. (2003), where a concise description of the “standard star-formation model” can be found. The observed nearby stellar luminosity function, Ψ_{near} , which is not corrected for Lutz–Kelker bias (Lutz and Kelker 1973, Tables 2 and 8 in Kroupa 1995a) and which is smoothed by using larger bin widths at the faint end, as detailed in Sect. 4 of that paper, is plotted as the solid-line histogram. The filled circles represent the best-estimate Malmquist corrected photometric luminosity function, $\bar{\Psi}_{\text{phot}}$ (☛ Fig. 4-9). By correcting for Malmquist bias (Stobie et al. 1989), the LF becomes that of a single-age, single-metallicity population (Taken from Kroupa (1995e))

of faint cluster members is very arduous because of contamination by the background or foreground Galactic-field population. The first step is to obtain photometry of everything stellar in the vicinity of a cluster and to select only those stars that lie near one isochrone, taking into account that unresolved binaries are brighter than single stars. The next step is to measure proper motions and radial velocities of all candidates to select only those high-probability members that have coinciding space motion with a dispersion consistent with the a priori unknown but estimated internal kinematics of the cluster. Since nearby clusters for which proper-motion

measurements are possible appear large on the sky, the observational effort is overwhelming. An excellent example of such work is the 3D mapping of the Hyades cluster by Röser et al. (2011). For clusters such as globulars that are isolated, the second step can be omitted, but in dense clusters, stars missed due to crowding need to be corrected for.

The stellar LFs in clusters turn out to have the same general shape as Ψ_{phot} (► Fig. 4-10), with the maximum being slightly offset depending on the metallicity of the population (► Figs. 4-13 and ► 4-14). A 100 Myr isochrone (the age of the Pleiades) is also plotted in ► Fig. 4-11 to emphasize that for young clusters additional structure (in this case, another maximum near $M_V = 8$ in the LF is expected via (► 4.1)). This is verified for the Pleiades cluster (Belikov et al. 1998) and is due to stars with $m < 0.6 M_{\odot}$ not having reached the main sequence yet (Chabrier and Baraffe 2000).

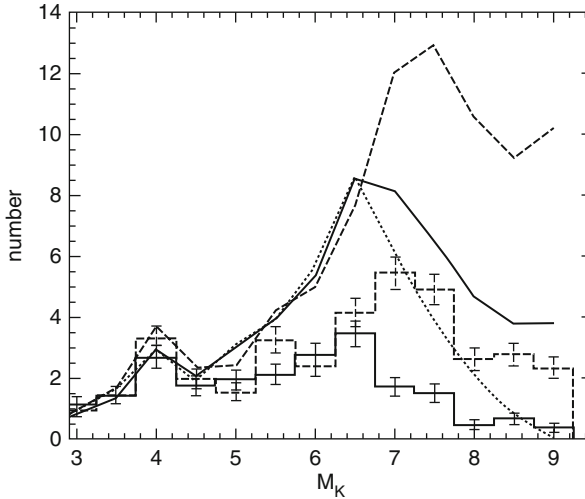
LFs for star clusters are, like Ψ_{phot} , system LFs because binary systems are not resolved in the typical star-count survey. The binary-star population evolves due to encounters, and after a few initial crossing times, only those binary systems survive that have a binding energy larger than the typical kinetic energy of stars in the cluster (Heggie 1975; Marks et al. 2011).

A further complication with cluster LFs is that star clusters preferentially lose single low-mass stars across the tidal boundary as a result of ever-continuing redistribution of energy during encounters while the retained population has an increasing binary proportion and increasing average stellar mass. The global PDMF thus flattens with time with a rate proportional to the fraction of the cluster lifetime, and for highly evolved initially rich open clusters, it evolves toward a delta function near the turnoff mass, the mass-loss rate being a function of galactocentric distance. This is a major issue for aged open clusters (initially $N < 10^4$ stars) with lifetimes of only a few 100 Myr.

These processes are now well quantified, and ► Fig. 4-16 shows that a dynamically very evolved cluster such as the Hyades has been depleted significantly in low-mass stars. Even so, the binary-star correction that needs to be applied to the LF in order to arrive at the individual-star present-day LF is significant.

A computationally challenging investigation of the systematic changes of the MF in evolving clusters of different masses has been published by Baumgardt and Makino (2003). Baumgardt and Makino quantify the depletion of the clusters of low-mass stars through energy-equipartition-driven evaporation and conclusively show that highly evolved clusters have a very substantially skewed PDMF (► Fig. 4-17). If the cluster ages are expressed in fractions, τ_f , of the overall cluster lifetime, which depends on the initial cluster mass, its concentration, and orbit, then different clusters on different orbits lead to virtually the same PDMFs at the same τ_f . Their results were obtained for clusters that are initially in dynamical equilibrium and that do not contain binary stars (these are computationally highly demanding), so that future analysis, including initially non-virialized clusters and a high primordial binary fraction (► Sect. 2.6), will be required to further refine these results.

For the massive and long-lived globular clusters ($N \gtrsim 10^5$ stars), theoretical stellar-dynamical work shows that the MF measured for stars near the cluster's half-mass radius is approximately similar to the global PDMF, while inward and outward of this radius the MF is flatter (smaller α_1) and steeper (larger α_1), respectively, owing to dynamical mass segregation (Vesperini and Heggie 1997). However, mass loss from the cluster flattens the global PDMF such that it no longer resembles the IMF anywhere (► Fig. 4-17), for which evidence has been found in some cases (Piotto and Zoccali 1999, see also ► Sect. 12.7). The MFs measured for globular clusters must therefore generally be flatter than the IMF, which is indeed born-out by observations (► Fig. 4-26 below). However, again the story is by no means straightforward because



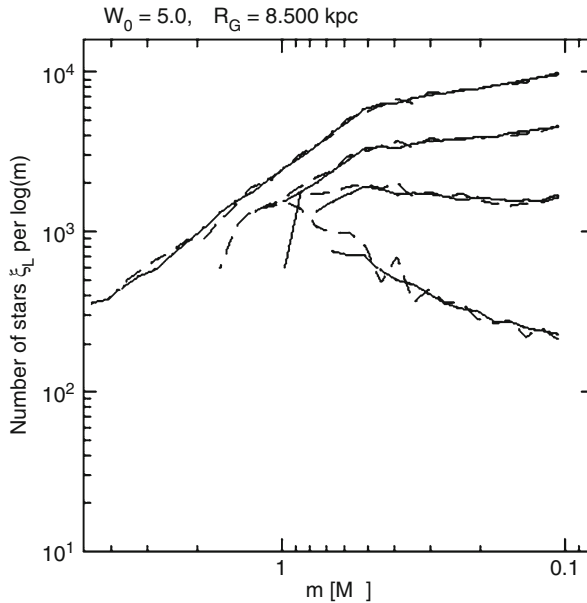
■ Fig. 4-16

Models of the K -band individual-star and system LFs in an ensemble of 20 dynamically highly evolved clusters (*thin* and *thick histograms*, respectively). An observer would deduce the *thick histogram*, which can only be transformed to the individual-star PDMF in the cluster if adequate correction for unresolved binaries is made. Such a correction leads to the *upper thin histogram* from which the PDMF can be inferred via (4.1). Each cluster model consists initially of 200 binaries with a half-mass radius of 0.85 pc, and the LFs are shown at an age of 480 Myr (44 initial crossing times) and count all stars and systems within the central sphere with a radius of 2 pc. The clusters are random renditions from the same parent distributions (binary-star orbital parameters, IMF, stellar positions, and velocities) and are initially in dynamical equilibrium. The *upper dashed curve* is the initial individual-star LF (KTG93 MLR, Fig. 4-11, and canonical IMF, 4.55) below), and the *solid curve* is the model Galactic-field LF of systems, also shown in Fig. 4-15. This is an accurate representation of the Galactic-field population in terms of the IMF and mixture of single and binary stars and is derived by stars forming in clusters such as shown here that dissolve with time. Both of these LFs are identical to the ones shown in Fig. 4-15. The *dotted curve* is the initial system LF (100% binaries) (From Kroupa (1995c))

globular clusters have significantly smaller binary fractions than population II clusters (Ivanova et al. 2005). The binary-star corrections are therefore smaller for globular cluster MFs implying a larger difference between α_1 for GCs and open clusters for which the binary correction is very significant.

Therefore, and as already pointed out by Kroupa (2001b), it appears quite realistically possible that population II IMFs were in fact flatter (smaller α_1) than population I IMFs, as would be qualitatively expected from simple fragmentation theory (Sect. 12.2). Clearly, this issue needs detailed investigation which, however, is computationally highly demanding, requiring the use of state-of-the-art N -body codes and special-purpose hardware.

The first realistic calculations of the formation of an open star cluster such as the Pleiades demonstrate that the binary properties of stars remaining in the cluster are comparable to those observed even if all stars initially form in binary systems according to the BBP ((4.46),

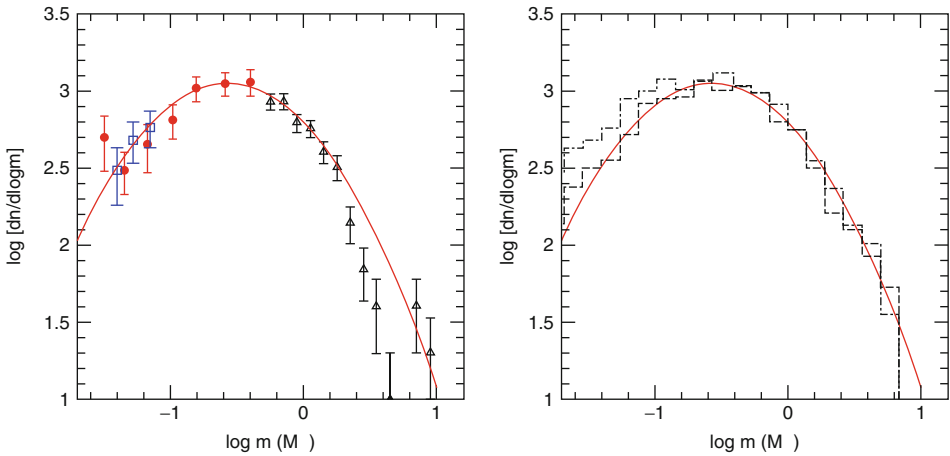


■ Fig. 4-17

PDMFs in a King-model cluster with concentration $W_0 = 5$ on a circular orbit about the MW center with radius 8.5 kpc. Shown are the MFs of all bound stars at ages corresponding to $\tau_f = 0\%$, 30%, 60%, and 90% of the cluster lifetime (from top to bottom). For each, age the *solid line* represents one computation with 1.28×10^5 stars, and the *dashed lines* show the sum of four clusters each with 8,000 stars (scaled to the same number of stars as the massive computation). Results for other circular and eccentric orbits and cluster concentrations are virtually indistinguishable (From Baumgardt and Makino (2003). Note the progressive depletion of low-mass stars as the cluster ages)

Kroupa et al. 2001, 2003). That work also demonstrates the complex and counterintuitive interplay between the initial concentration, mass segregation at the time of residual gas expulsion, and the final ratio of the number of BDs to stars (● Fig. 4-18). Thus, this modeling shows that an initially denser cluster evolves to significant mass segregation when the gas explosively leaves the system. Contrary to naive expectation, according to which a mass-segregated cluster should lose more of its least massive members during expansion after gas expulsion, the ensuing violent relaxation of the cluster retains more free-floating BDs than the less-dense model. This comes about because BDs are split from the stellar binaries more efficiently in the denser cluster. This, however, depends on the BDs and stars following the same pairing rules, which is now excluded (● Sect. 8). During the long-term evolution of the mass function, initially mass-segregated star clusters lose more low-mass stars when they are close to dissolution as has been found by Baumgardt et al. (2008) after comparing star loss from clusters in N -body calculations with the observed mass functions of globular clusters. But additionally, the gas expulsion from embedded clusters has to be carefully taken into account to explain the correlation between concentration of a globular cluster and the slope of its PDMF (● Sect. 12.7).

These issues remain an active area of research because at least two changes need to be made to the modeling: On the one hand, BDs need to be treated as a population separate from the



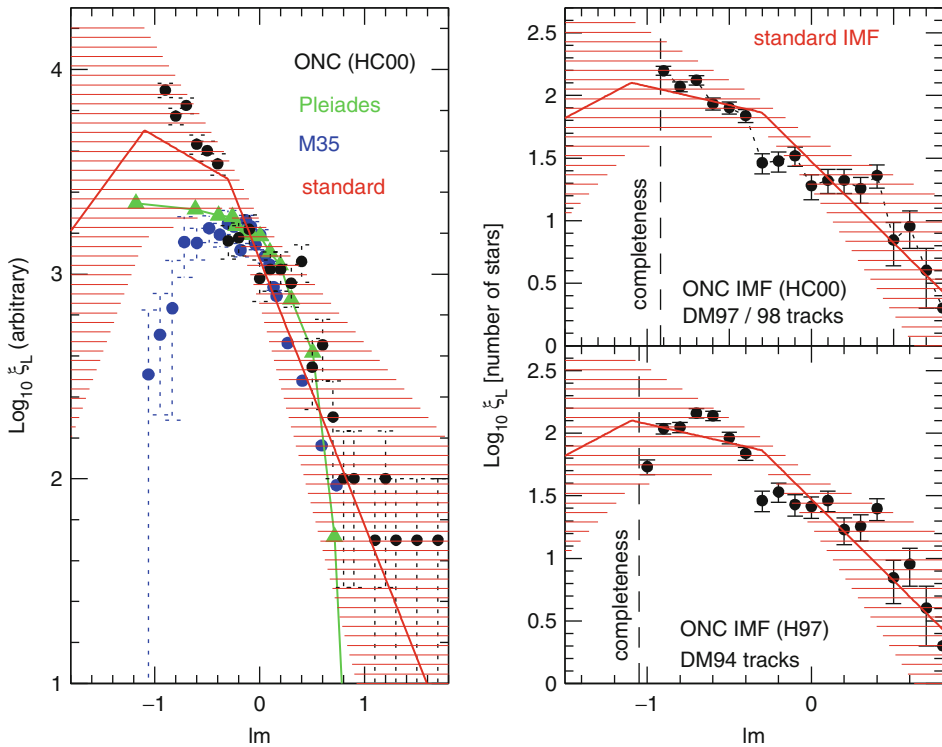
■ Fig. 4-18

The observationally deduced MF in the Pleiades cluster. *Left panel:* The symbols are observational data (for details, see Moraux et al. 2004) and the curve is a log-normal fit. *Right panel:* The curve is the same log-normal fit. Theoretical system MFs for two initial models of the Pleiades cluster according to Kroupa et al. (2001) are plotted at an age of 100 Myr. These models assume the young cluster to be deeply embedded in gas, with a star-formation efficiency of 33%, a gas-expulsion time scale shorter than the crossing time and to contain 10^4 stars and BDs, whereby all stars and BDs are paired randomly to binary systems (i.e., in these models, BDs are not treated as a separate population, see Sect. 8 for more realistic models). Model A (*dashed histogram*) has an initial central number density $\rho_C = 10^{4.8}$ stars/pc³, while model B (*dotted histogram*) has $\rho_C = 10^{5.8}$ stars/pc³. The embedded phase lasts 0.6 Myr, and during this time, mass segregation develops in the initially denser cluster model B. Note that these models are not a fit but a prediction of the Pleiades MF, assuming it had a canonical IMF (4.55). Note that the initially denser cluster (*upper histogram*) retains more BDs as a result of these being ionized off their stellar primaries in the denser environment of model B. Also, note that the observational data suggest a deficit of early-type stars in the Pleiades (*left panel*) which is reminiscent of the deficit of massive stars noted for the ONC (Pflamm-Altenburg and Kroupa 2006)

stellar one (Sect. 8) so that the free-floating BDs that result, in the currently available models, from the disruption of star-BD binaries, will not be available in reality. On the other hand, some observations suggest that star clusters may form highly mass-segregated. The mass-dependent loss of stars thus definitely remains an issue to be studied.

The above work suggests that even clusters as young as the Pleiades are significantly evolved because clusters of all masses form from highly concentrated embedded morphologies (Kroupa 2005; Marks and Kroupa 2010, 2012; Conroy 2011). Also, the low-mass stars in clusters as young as the Pleiades or M35 (Fig. 4-19 below) have not yet reached the main sequence, so that pre-main sequence stellar-evolution calculations have to be resorted to when transforming measured luminosities to stellar masses via the MLR.

For ages younger than a few Myr, this becomes a serious problem: Classical pre-main sequence theory, which assumes hydrostatic contraction of spherical non-, sometimes

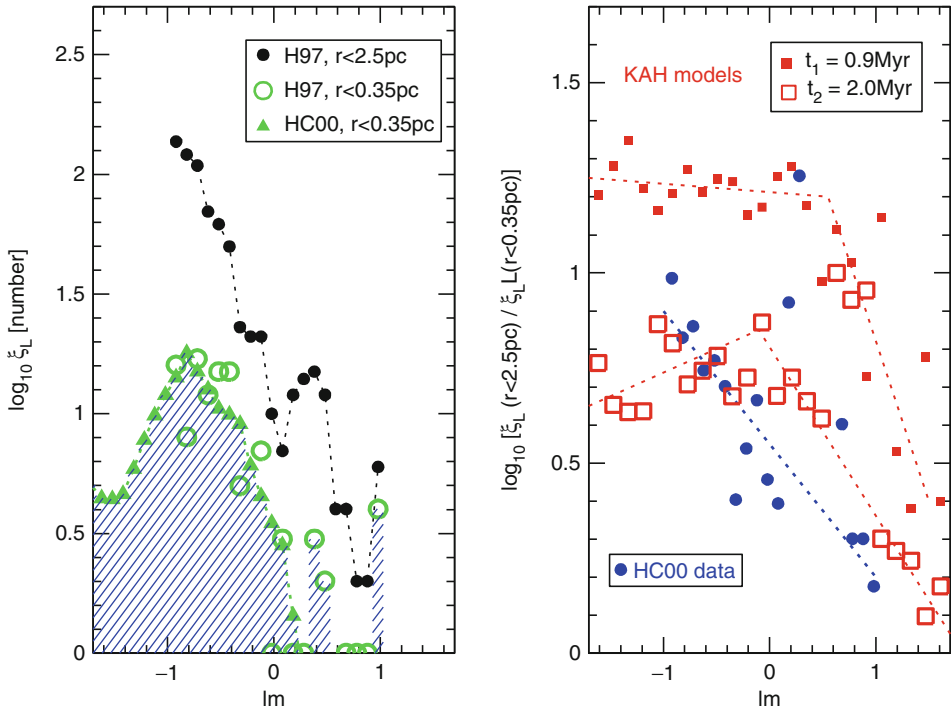


■ Fig. 4-19

Left panel: The observationally deduced system mass functions ($\text{lm} \equiv \log_{10}(m/M_{\odot})$) (a) in the ONC (Hillenbrand and Carpenter 2000): optical data within $r \leq 2.5$ pc, $\tau_{\text{cl}} < 2$ Myr, $[\text{Fe}/\text{H}] = -0.02$ (Esteban et al. 1998) (b) in the Pleiades (Hambly et al. 1999): $r \leq 6.7$ pc, $\tau_{\text{cl}} \approx 100$ Myr, $[\text{Fe}/\text{H}] = +0.01$, and (c) in M35 (Barrado y Navascués et al. 2001): $r \leq 4.1$ pc, $\tau_{\text{cl}} \approx 160$ Myr, $[\text{Fe}/\text{H}] = -0.21$, where r is the approximate projected radius of the survey around the cluster center and τ_{cl} the nuclear age. The strong decrease of the M35 MF below $m \approx 0.5 M_{\odot}$ remains present despite using different MLRs (e.g., DM97, as in the *right panel*). None of these MFs are corrected for unresolved binary systems. The canonical individual-star IMF (4.55) is plotted as the three *straight lines* assuming here continuity across the VLMS/BD mass range. *Right panel:* The shape of the ONC IMF differs significantly for $m < 0.22 M_{\odot}$ if different pre-main sequence evolution tracks, and thus essentially different theoretical MLRs, are employed (DM stands for tracks calculated by D’Antona and Mazzitelli, see Hillenbrand and Carpenter 2000 for details)

slowly-rotating stars from idealized initial states, breaks down because of the overlap with the star-formation processes that defies detailed treatment. Stars this young remember their accretion history, invalidating the application of classical pre-main sequence stellar evolution tracks, a point made explicitly clear by Tout et al. (1999), Wuchterl and Klessen (2001), and Wuchterl and Tscharnuter (2003), and are in any case rotating rapidly and are nonspherical. Such realistic pre-main sequence tracks are not available yet. The uncertainties due to such processes have been partially discussed though (Baraffe et al. 2002, 2009; Baraffe and Chabrier 2010).

Research on the IMF in very young clusters benefits from spectroscopic classification of individual stars to place them on a theoretical isochrone of existing classical pre-main-sequence



■ Fig. 4-20

Left panel: Mass segregation is very pronounced in the ONC, as is evident by comparing the observationally deduced MF for all stars within $r = 2.5$ pc with the MF for all stars with $r < 0.35$ pc (Hillenbrand and Carpenter 2000, HC00) ($l_m \equiv \log_{10}(m/M_{\odot})$). For both samples, the reddening $A_V < 2.5$ mag (Hillenbrand 1997, H97, is for an optical and spectroscopic survey, whereas HC00 is a near-infrared survey). *Right panel:* The ratio, $\xi_L(r < 2.5 \text{ pc}) / \xi_L(r < 0.35 \text{ pc})$ (solid dots), of the MFs shown in the left panel increases significantly with decreasing mass, demonstrating the significant depletion of low-mass stars in the central region of the ONC. Stellar-dynamical models of the ONC (Kroupa et al. 2001) approximately reproduce this trend at an age of 2 Myr for the canonical IMF (☛ 4.55), whereby the system masses of surviving binary systems are counted instead of the individual stars many of which are in unresolved binaries even if no initial mass segregation is assumed (at $t = 0$, $\xi_L(r < 2.5 \text{ pc}) / \xi_L(r < 0.35 \text{ pc}) = \text{constant}$). The model snapshots shown are from model B in Kroupa et al. (2001) under the assumption that prior to gas-expulsion, the central stellar density was $\rho_c = 10^{5.8}$ stars/pc³. The dotted lines are eye-ball fits to the plotted data (See also ☛ Fig. 4-18)

evolution theory to estimate masses (e.g., Meyer et al. 2000; Luhman 2004; Barrado y Navascués et al. 2004; Slesnick et al. 2004). In such cases, the deduced age spread becomes comparable to the age of the cluster (☛ 4.3). Binary systems are mostly not resolved but can feign an apparent age spread even if there is none in the underlying population as has been shown by Weidner et al. (2009). Differential reddening due to inhomogeneously distributed remnant gas and dust has a significant effect on estimating stellar masses (Andersen et al. 2009). But also episodic accretion onto the protostars can mimic such age spreads (Tout et al. 1999; Baraffe et al. 2009; Baraffe and Chabrier 2010). The reality of age spreads is an important issue as it defines whether

star clusters are almost coeval or host prolonged star formation. The finding of 10–30 Myr old dwarfs by the Lithium depletion method in the ≈ 1 Myr ONC by Palla et al. (2007) seems to indicate the latter. But careful numerical calculations have shown that a collapsing molecular cloud can trap a corresponding amount of stars from a surrounding older OB association into the forming cluster (Pflamm-Altenburg and Kroupa 2007). The ONC is embedded in the Orion OB1 association that has an age of 10–15 Myr which could explain the older dwarfs in the ONC.

A few results are shown in [Figs. 4-19](#) and [4-20](#). While the usual argument is for an invariant IMF, as is apparent for most population I stars (e.g., Fig. 5 in Chabrier 2003a; Fig. 3 in Bastian et al. 2010), [Fig. 4-19](#) shows that some appreciable differences in measured MFs are evident. The M35 MF appears to be highly deficient in low-mass stars. This clearly needs further study because M35 and the Pleiades appear to be otherwise fairly similar in terms of age, metallicity (M35 is somewhat less metal-rich than the Pleiades), and the size of the survey volume.

Taking the ONC as the best-studied example of a very young and nearby rich cluster (age ≈ 1 Myr, distance ≈ 450 pc; $N \approx 5,000$ –10,000 stars and BDs; Hillenbrand and Carpenter 2000; Luhman et al. 2000; Muench et al. 2000; Kroupa 2000; Slesnick et al. 2004), [Fig. 4-19](#) shows how the shape of the deduced IMF varies with improving (but still classical) pre-main sequence contraction tracks. This demonstrates that any substructure cannot, at present, be relied upon to reflect possible underlying physical mechanisms of star formation.

Main Results

Currently available evidence from resolved stellar populations largely points toward an IMF of late-type ($m \lesssim 1 M_{\odot}$) stars which is independent of the environment and which can be described well by a power law with an index of about $\alpha_2 = 2.3$ for $m \gtrsim 0.5 M_{\odot}$ and $\alpha_1 = 1.3$ for $m \lesssim 0.5 M_{\odot}$.

8 The IMF of Very Low-Mass Stars (VLMSs) and of Brown Dwarfs (BDs)

These are stars near to the hydrogen-burning mass limit (VLMS) or objects below it (BDs). BDs are not massive enough to achieve sufficiently high central pressures and temperatures to stabilize against continued contraction by burning H and thus indefinitely cool to unobservable luminosities and temperatures. The term “brown dwarf” was coined by Jill Tarter in her 1975 Ph.D. thesis and was later generally accepted as a name for substellar (i.e., non-hydrogen-burning) objects which presumably form like stars.

Observationally, it is very difficult to distinguish between VLMSs and BDs because a sufficiently young BD may have colors and spectral features corresponding to a VLMS. BDs were studied as theoretical objects in 1963 by Hayashi and Nakano (1963), who performed the first truly self-consistent estimate of the minimum hydrogen burning mass limit, m_H , by computing the luminosity at the surface and the energy release rate by nuclear burning. Modern theory of the evolution and internal constitution of BDs has advanced considerably owing to the inclusion of an improved equation of state and realistic model atmospheres that take into account absorption by many molecular species as well as dust allowing the identification of characteristic photometric signatures (Chabrier and Baraffe 2000). This work shows that the critical mass below which an object cannot be stabilized by nuclear fusion is $m_H = 0.075 M_{\odot}$ for solar metallicity. For lower metallicity, m_H is larger since a larger luminosity (due to the lower opacity)

requires more efficient nuclear burning to reach thermal equilibrium and thus a larger mass. The first BDs were detected in 1995, and since then, they have been found in the solar neighborhood and in young star clusters (Basri 2000), allowing increasingly sophisticated estimates of their mass distribution (Bouvier et al. 2003).

For the solar neighborhood, near-infrared large-scale surveys have now identified many dozens of BDs probably closer than 25 pc (e.g., Allen et al. 2005). Since these objects do not have reliable distance measurements, an ambiguity exists between their ages and distances, and only statistical analysis that relies on an assumed star-formation history for the solar neighborhood can presently constrain the IMF (Chabrier 2002), finding a 60% confidence interval $\alpha_0 = 0.3 \pm 0.6$ for 0.04–0.08 M_\odot approximately for the Galactic-field BD IMF (Allen et al. 2005).

Surveys of young star clusters have also discovered BDs by finding objects that extend the color–magnitude relation toward the faint locus while being kinematical members. Given the great difficulty of this endeavor, only a few clusters now possess constraints on the MF. The Pleiades star cluster has proven especially useful, given its proximity ($d \approx 127$ pc) and young age ($\tau_{\text{cl}} \approx 100$ Myr). Results indicate $\alpha_0 \approx 0.5$ –0.6. Estimates for other clusters (ONC, σ Ori, IC 348, Cha I) also indicate $\alpha_0 \lesssim 0.8$. In their Table 1, Allen et al. (2005) summarize the available measurements for 11 populations finding that $\alpha_0 \approx 0$ –1 and Andersen et al. (2008) find the low-mass IMF in seven young star-forming regions to be most consistent with being sampled from an underlying log-normal (Chabrier) IMF.

However, while the log-normal (Chabrier) IMF is indistinguishable in the stellar regime from the simpler canonical two-part power-law IMF (see [Fig. 4-24](#) below), it is to be noted that these and other constraints on the IMF of BDs rely on assuming the IMF to be continuous across the stellar/BD boundary. In the following, it will emerge that this assumption is not consistent with the binary properties of stars and BDs. The IMF can therefore not be a continuous log-normal across the VLMS/BD boundary.

8.1 BD and VLMS Binaries

The above estimates of the BD IMF suffer under the same bias affecting stars, namely, from unseen companions which need to be taken into account to infer the true BD IMF.

BD–BD binary systems are known to exist (Basri 2000) in the field (Bouy et al. 2003; Close et al. 2003) and in clusters (Martín et al. 2003). Their frequency is not yet fully constrained since detailed scrutiny of individual objects is time-intensive on large telescopes, but the data suggest a binary fraction of about 15% only. The results show conclusively that the semimajor axis distribution of VLMSs and BDs is much more compact than that of M dwarfs, K dwarfs, and G dwarfs. Bouy et al. (2003), Close et al. (2003), Martín et al. (2003), and Phan-Bao et al. (2005) all find that BD binaries with semimajor axis $a \gtrsim 15$ AU are very rare. Using Monte-Carlo experiments on published multiple-epoch radial-velocity data of VLMSs and BDs, Maxted and Jeffries (2005) deduce an overall binary fraction between 32% and 45% with a semimajor axis distribution that peaks near 4 AU and is truncated at about 20 AU. In the Pleiades cluster where their offset in the color–magnitude diagram from the single-BD locus makes them conspicuous, Pinfield et al. (2003) find the BD binary fraction may be as high as 60%. This, however, appears unlikely as a survey of more than 6 years UVES/VLT spectroscopy has shown the BD binary fraction to be 10–30% and that incompleteness in the few AU separation region is not significant (Joergens 2008). Using a very deep infrared survey of the Pleiades, Lodieu et al. (2007) suggest a BD binary fraction of 28 to 44 per cent, consistent with the Maxted and Jeffries (2005) result but only marginally so with the Pinfield et al. (2003) value.

It has already been shown that the disruption in embedded clusters of stellar binaries in a stellar population which initially consists of 100% binary stars with periods (i.e., binding energies) consistent with the observed pre-main sequence and proto-stellar binary data (☉ 4.46) leads to the observed main-sequence binary population in the Galactic field (Kroupa 1995d; Goodwin and Kroupa 2005; Marks and Kroupa 2011). Systems with BD companions have an even lower binding energy, and the truncated semimajor axis distribution of BDs may be a result of binary disruption in dense clusters of an initial stellar-like distribution.

This notion is tested by setting up the *star-like hypothesis* (Kroupa et al. 2003):

Star-like Hypothesis for BDs

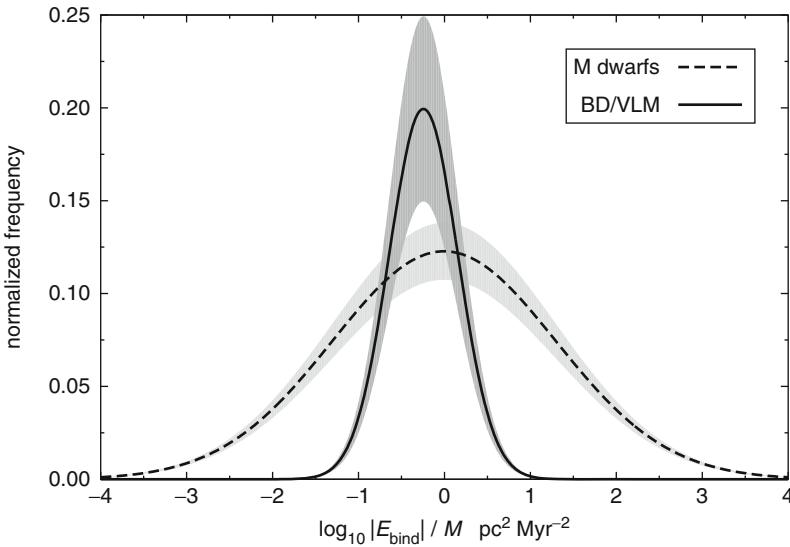
BDs form as stars do from molecular-cloud cores.

The star-like hypothesis implies that BDs form according to the same binary pairing rules (the BBP, ☉ 4.46) as stars do.

If BDs form as stars do then this hypothesis ought to be true since objects with masses $0.04\text{--}0.07 M_{\odot}$ should not have very different pairing rules than stars that span a much larger range of masses ($0.1\text{--}1 M_{\odot}$) but show virtually the same period-distribution function independently of primary mass (the M-, K-, and G-dwarf samples, Fischer and Marcy 1992; Mayor et al. 1992; Duquennoy and Mayor 1991, respectively). Thus, the hypothesis is motivated by observed orbital distribution functions of stellar binaries not being sensitive to the primary mass, which must come about if the overall physics of the formation problem is similar. Further arguments for a star-like origin of BDs comes from the detection of accretion onto and disks about very young BDs and that the BDs and stars in Taurus-Auriga have indistinguishable spatial and velocity distributions (White and Basri 2003).

Assuming the star-like hypothesis to hold, Kroupa et al. (2001, 2003) perform N -body calculations of ONC- and Taurus-Auriga-like stellar aggregates to predict the semimajor axis distribution functions of BD–BD, star–BD, and star–star binaries. These calculations demonstrate that the binary proportion among BDs becomes smaller than among low-mass stars after a few crossing times, owing to their weaker binding energies. The distribution of separations, however, extends to similar distances as for stellar systems (up to $a \approx 10^3$ AU), disagreeing completely with the observed BD–BD binary distribution. The star-like hypothesis thus predicts far too many wide BD–BD binaries. This can also be seen from the distribution of binding energies of real BD binaries. It is very different to that of stars by having a low-energy cutoff, ${}^{\text{BD}}E_{\text{bin,cut}} \approx -10^{-0.9} M_{\odot} (\text{pc/Myr})^2$, that is much higher than that of the M dwarfs, ${}^{\text{M}}E_{\text{bin,cut}} \approx -10^{-3} M_{\odot} (\text{pc/Myr})^2$ (☉ Fig. 4-21). This is a very strong indicator for some fundamental difference in the dynamical history of BDs.

Furthermore, the N -body distributions contain a substantial number of star–BD pairs, which also disagrees with the existence of very few BD companions to nearby stars (Basri 2000; Phan-Bao et al. 2005). Basically, if BDs form exactly like stars, then the number of star–BD binaries would be significantly larger than is observed, since, for example, G-dwarfs prefer to pair with M-dwarfs (why should BDs be any different from M dwarfs in their pairing to G dwarfs?). The observed general absence of BD companions is referred to as the *BD desert* (Zucker and Mazeh 2001) since stellar companions and planets are found at such separations (Halbwachs et al. 2000; Vogt et al. 2002). A few very wide star–BD systems can form during the final stages of dissolution of a small cluster (de La Fuente Marcos 1998), and three such common proper-motion pairs have perhaps been found (Gizis et al. 2001).



■ Fig. 4-21

The distribution of binding energies, $E_{\text{bin}} = -G m_1 m_2 / (2 a)$, of BDs (solid line) compared to those of M dwarfs (MDs, dashed line). The BD distribution is computed as a Gaussian distribution based on BD/VLMs data from the Very Low Mass Binary Archive (<http://vlmbinaries.org>). Specifically, the Gaussian distribution has a mean semimajor axis $\log_{10}(a_{\text{mean}}/\text{AU}) = 0.6$ with a half width of $\log_{10}(\sigma_a/\text{AU}) = 0.4$. The *upper* and *lower* envelopes correspond to BD binary fractions of $f_{\text{BD}} = 0.25$ and 0.15 , respectively (area under the curves). The MD energy distribution is computed by assuming the a -distribution from Fischer and Marcy (1992) (which is practically identical to that of G dwarfs) and choosing 10^7 masses, $m_i \in (0.1 - 0.5 M_{\odot})$, from the canonical stellar IMF (4.55) and random pairing. For the MDs, the Gaussian distribution has $\log_{10}(a_{\text{mean}}/\text{AU}) = 1.5$ and $\log_{10}(\sigma_a/\text{AU}) = 1.3$ such that $f_{\text{MD}} = 0.45$ and 0.35 for the *upper* and *lower* envelopes, respectively

Finally, the star-like hypothesis also predicts far too few star–star binaries in Taurus–Auriga, where binary disruption has not been active. This comes about because the large number of star–BD systems in this model limits the number of star–star binaries given the finite number of stellar primaries.

It is thus concluded that the observed BD population is incompatible with the star-like hypothesis. Therefore, BDs need to be treated as a separate, or extra, population (Kroupa et al. 2003). This is confirmed by Parker and Goodwin (2011), who constrain BD binary properties from the observed ones given that their dynamical evolution in the birth clusters needs to be corrected for.

BD/Star Population Synthesis Theorem

When setting up a population of BDs and stars, BDs need to be algorithmically treated separately from the stars.

Proof: The BD desert, see the rejection of the star-like hypothesis above and/or Kroupa et al. (2003). A practical formulation of this theorem is posed as a Gedanken Experiment on p. 184. \square

8.2 The Number of BDs per Star and BD Universality

Briceño et al. (2002) report that Taurus-Auriga appears to form significantly fewer BDs per star than the ONC. Both systems are very different physically but have similar ages of about 1 Myr. This finding was interpreted to be the first possible direct evidence of a variable IMF, being consistent qualitatively with the Jean-mass,

$$M_J \propto \rho^{-1/2} T^{3/2}, \quad (4.50)$$

being larger in Taurus-Auriga than in the ONC because its gas density, ρ , is smaller by one to two orders of magnitude, while the temperatures, T , are similar to within a factor of a few (\blacktriangleright Sect. 1.4).

Given this potentially important finding, Kroupa et al. (2003) computed N -body models of the stellar aggregates in Taurus-Auriga in order to investigate the hypothesis that BDs form star-like. They find that the same initial number of BDs per star in Taurus-Auriga and in the ONC leads to different observed ratios because BD–BD and star–BD binaries are disrupted more efficiently in the ONC; the observer thus sees many more BDs there than in the comparatively dynamically unevolved Taurus-Auriga groups. But, as already noted above, the star-like hypothesis must be discarded because it leads to too many wide BD–BD binaries, and also it predicts too many star–BD binaries. Given this problem, Kroupa and Bouvier (2003a) study the production rate of BDs per star assuming BDs are a separate population, such as ejected embryos (Reipurth and Clarke 2001), or as gravitational instabilities in extended circum-proto-stellar disks (Goodwin and Whitworth 2007; Thies et al. 2010). Again, they find that both, the physically very different environments of Taurus-Auriga and the ONC, can have produced the same ratios (about one BD per four stars) if BDs are ejected embryos with a dispersion of ejection velocities of about 2 km/s (this number is revised to 1.3 km/s below).

Based on some additional observations, Luhman (2004) revised the Briceño et al. (2002) results by finding that the number of BDs per star had been underestimated in Taurus-Auriga. Since the new spectroscopic study of Slesnick et al. (2004) also revised the number of BDs per star in the ONC downward, Luhman (2004) retracts the significance of the claimed difference of the ratio in Taurus-Auriga and the ONC. Is a universal, invariant, BD production scenario still consistent with the updated numbers?

Let the true ratio of the number of BDs per late-type star be

$$R \equiv \frac{N(0.02 - 0.08 M_{\odot})}{N(0.15 - 1.0 M_{\odot})} \equiv \frac{N_{\text{BD,tot}}}{N_{\text{st,tot}}}. \quad (4.51)$$

Note that here stars more massive than $1.0 M_{\odot}$ are not counted because Taurus-Auriga is mostly producing late-type stars given the limited gas mass available (see also \blacktriangleright Fig. 4-5). But the observed ratio is

$$R_{\text{obs}} = \frac{N_{\text{BD,obs}}}{N_{\text{st,obs}}} = N_{\text{BD,tot}} (\mathcal{B} + \mathcal{U}) \frac{(1+f)}{N_{\text{st,tot}}} = R (\mathcal{B} + \mathcal{U}) (1+f), \quad (4.52)$$

since the observed number of BDs, $N_{\text{BD,obs}}$, is the total number produced multiplied by the fraction of BDs that are gravitationally bound to the population (\mathcal{B}) plus the unbound fraction,

\mathcal{U} , which did not yet have enough time to leave the survey area. These fractions can be computed for dynamical models of the Taurus-Auriga and ONC and depend on the mass of the Taurus-Auriga subgroups and of the ONC and on the dispersion of velocity of the BDs. This velocity dispersion can either be the same as that of the stars if BDs form like stars or larger if they are ejected embryos (Reipurth and Clarke 2001). The observed number of “stars” is actually the number of systems such that the total number of individual stars is $N_{\text{st,tot}} = (1 + f) N_{\text{st,obs}}$, where f is the binary fraction of stars (► 4.48). Note that here, no distinction is made between single or binary BDs, which is reasonable given the low binary fraction (about 15%) of BDs. For Taurus-Auriga, Luhman (2004) observes $R_{\text{TA,obs}} = 0.25$ from which follows

$$R_{\text{TA}} = 0.18 \text{ since } f_{\text{TA}} = 1, \mathcal{B} + \mathcal{U} = 0.35 + 0.35 \quad (4.53)$$

(Kroupa et al. 2003). According to Slesnick et al. (2004), the revised ratio for the ONC is $R_{\text{ONC,obs}} = 0.28$ so that

$$R_{\text{ONC}} = 0.19 \text{ because } f_{\text{ONC}} = 0.5, \mathcal{B} + \mathcal{U} = 1 + 0 \quad (4.54)$$

(Kroupa et al. 2003). Note that the regions around the stellar groupings in Taurus-Auriga not yet surveyed should contain about 30% of all BDs, while all BDs are retained in the ONC. The ONC and TA thus appear to be producing quite comparable BD/star number ratios.

Therefore, the updated numbers imply that about *one BD is produced per five late-type stars* and that the dispersion of ejection velocities is $\sigma_{\text{ej}} \approx 1.3$ km/s. These numbers are an update of those given in Kroupa et al. (2003), but the results have not changed much. Note that a BD with a mass of $0.06 M_{\odot}$ and a velocity of 1.3 km/s has a kinetic energy of $10^{-1.29} M_{\odot} (\text{pc/Myr})^2$ which is rather comparable to the cutoff in BD–BD binding energies (► Fig. 4-21). *This supports the notion that most BDs may be mildly ejected embryos*, for example, from the Goodwin and Whitworth (2007) and Thies et al. (2010) circum-*proto-stellar* disks.

It appears thus that the different physical environments evident in Taurus-Auriga and the ONC produce about the same number of BDs per late-type star, so that there is no convincing evidence for differences in the IMF among current nearby star-forming regions across the hydrogen burning mass limit.

There is also no substantial evidence for a difference in the *stellar* IMF in these two star-forming regions, contrary to the assertion by, for example, Luhman (2004). ► Figure 4-22 shows the MF in four young clusters. The caveat that the classical pre-main sequence evolution tracks, upon which the observational mass determinations rely, are really applicable for such young ages (► Sect. 7.4) needs to be kept in mind though. Also, the observationally derived MFs are typically obtained by treating the luminous objects as single, while a majority are likely to be binary.

8.3 BD Flavors

BDs can come in different flavors depending on their formation (Kroupa and Bouvier 2003a): star-like BDs, ejected embryos, collisional BDs, and photo-evaporated BDs. As seen above, star-like BDs appear to be very rare because BDs do not mix with stars in terms of pairing properties. The rarity of the star-like BD flavour is supported theoretically because in order to have a cloud core produce only a BD or a BD-BD binary it needs to acquire an extreme density

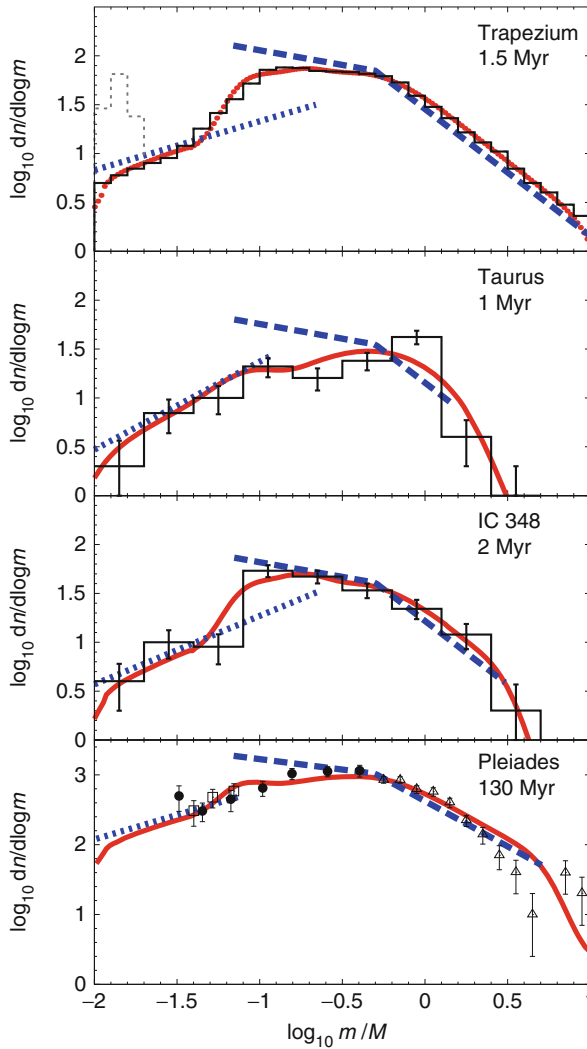


Fig. 4-22

The observationally determined MFs in four young clusters, the names and approximate ages of which are indicated in the panels, are shown as the histograms or data with error bars. The *thick solid (red) curve* is the model of unresolved systems by Thies and Kroupa (2007). In this model, the stars have a binary fraction of 100% in Taurus and 50% in the other cases, with companion masses selected randomly from the canonical IMF (4.46). The two-part power-law canonical IMF (4.55) is shown as the *thick dashed line*. The BDs are described by the additional power-law IMF shown as the *dotted (blue) line*. In the model, they have a binary fraction of 15% and do not mix with the stars (the brown dwarf desert). It can be seen that the model is essentially invariant and leads to an excellent description of the data despite the BD and stellar IMFs being discontinuous (4.23). Further details are found in Sect. 8.4

to become gravitationally unstable towards collapse. Such conditions are very rare in typical turbulent molecular clouds that have Mach numbers less than about 6 (Hennebelle and Chabrier 2008, 2009).

A more recently discussed possible channel of BD formation is via the fragmentation of massive circum-stellar disks (FMCSs) beyond about 100 AU and subsequent dynamical separation of the resulting weakly bound BD companion (Goodwin and Whitworth 2007; Stamatellos et al. 2007a; Stamatellos and Whitworth 2009; Thies et al. 2010; Basu and Vorobyov 2012). This mechanism for producing BDs can be considered a revised version of the ejected BD scenario. It appears to be the most promising physical mechanism for producing the dominant flavor of BDs.

An environmental dependency of the VLMS and BD IMF is expected for the FMCS channel as shown by Stamatellos et al. (2011). They point out that disk fragmentation is enhanced when the star plus disk accretes episodically as long as the accretion intervals are long enough to allow the disk to cool. In dense environments, episodic accretion would occur too frequently thereby inhibiting fragmentation as the accreting star's luminosity is kept as a high accretion luminosity thereby heating the disk.

The collisional removal of accretion envelopes for the production of unfinished stars needs to be discounted as a source for BDs because this process is far too rare (Kroupa and Bouvier 2003a). The removal of accretion envelopes through photo-evaporation can occur, but only within the immediate vicinity of an O star and never in Taurus-Auriga (Whitworth and Zinnecker 2004). Even in the presence of some ionising stars, as in the cluster NGC 6611 which ionises the Eagle nebula, Oliveira et al. (2009) find no measurable effect of photo-evaporation on the sub-stellar MF. However, Kroupa and Bouvier (2003a) show that the radius within which photo-evaporation may be able to remove substantial fractions of an accretion envelope within 10^5 year is comparable to the cluster size in star-burst clusters that contain thousands of O stars. In such clusters, photo-evaporated BDs may be very common. Globular clusters (GCs) may then be full of BDs.

GC-BD Hypothesis

In GCs, the number ratio of BDs to stars may be very large ($\gg 1/5$).

8.4 The Origin of BDs and Their IMF

It has thus emerged that in order to construct a realistic stellar and BD population, BDs and VLMSs need to be treated according to a different initialization algorithm to that of stars. They need to be separated when setting up the binary populations. This can be visualized with the following Gedanken Experiment which is a practical formulation of the BD/Star Population Theorem on p. 180.

Gedanken Experiment

Imagine a box contains BDs, M-, K-, and G-dwarf stars. In order to pair these to obtain the correct birth binary population (p. 28), a distinction between M-, K-, and G-dwarfs need not be made. But, nearly every time a BD is picked as a companion to a star, or a star is picked as a companion to a BD, the system needs to be discarded.

The physical interpretation of this mathematical result is that BDs form along with stars, just as planets do, but, just like planets, they do not result from the same formation mechanism. Rather, similarly as planets, BDs stem from gravitationally preprocessed material.

Indeed, the conditions in a molecular cloud core very rarely are such that a dense-enough core can collapse under self-gravity without accreting too much material for it to not transcend the BD/star mass boundary, which is why most BDs do not derive from a star-like origin (see also [Sect. 11.2](#)). But the outer regions of circum-proto-stellar disks accumulate material which has time to lose entropy ([Sect. 8.3](#)). These outer regions are sufficiently dense to locally collapse under self-gravity either because they become unstable or because they are perturbed. They are not too dense and thus remain optically thin for a sufficiently long time to allow the collapsing object to cool radiatively, and the region around the collapsing object has a limited supply of local disk material. Fragmentation may occur in marginally stable disks upon an external perturbation by the gravity of other stars (Thies et al. 2005, 2010) or by the gas-dynamical interaction of two disks (Shen et al. 2010). Both such processes thus enlarge the parameter space of circum-stellar disk conditions for BD formation since most stars form in a clustered mode.

A circum-proto-stellar disk therefore sets the boundary conditions in favor of BD formation. This is the same, but even more extreme, for planet formation, for which a highly processed molecular cloud core is required in the sense that gravo-hydrodynamical dynamics is augmented significantly by the dynamics between solid particles in a circum-stellar disk. Thus, while the formation of BDs is still a purely gravo-hydrodynamical process within an extended disk, planet formation is seeded by the coagulation of dust to larger solids which may then induce gas accretion from the circum-stellar disk.

The differences in the binary properties of BDs and stars therefore indicate that they are two different classes of objects with their own separate mass distributions. The MF of planets is also never considered a continuous extension of the stellar IMF (Dominik 2011).

The IMF of BDs needs to be derived from the observational star count data by taking the above into account. Thies and Kroupa (2007, 2008) have done so. The results for Taurus-Auriga, the ONC, IC 348, and Pleiades demonstrate that the IMF appears to be universal with $\alpha_0 \approx 0.3$ albeit with a significant discontinuity near $0.1 M_\odot$ ([Fig. 4-22](#)). The data imply that there is an overlap region: For the analysis to correctly account for the observed data, there must be VLMSs that form as BDs do, while there are massive BDs that form as stars do ([Fig. 4-23](#)). Without account of this overlap the IMF as well as the binary properties as a function of the mass may feign continuity (as suggested by Kaplan et al. 2012). Noteworthy is that this analysis re-derives the same BD-to-star fraction as deduced above: About one BD forms per five stars. The same result on the number ratio has been inferred by Andersen et al. (2008) who, however, need to describe the BD MF as a decreasing log-normal form as a result of insisting the IMF to be continuous across the BD/stellar mass range. As stated above, this approach cannot account for the BD desert, and the correct approach is to treat the BDs/VLMSs as a distinct population from the stars implying two separate IMFs for BDs and stars. This correct approach leads to a power-law solution for the BD MF with $\alpha \approx 0.3$, that is, a mildly rising MF toward small masses, and separate BD and stellar IMFs.

The universality of the BD–stellar IMF is interesting and may suggest that the formation of BDs is mostly dependent on the conditions prevalent in circum-proto-stellar disks. Indeed, in a large ensemble of SPH simulations of circum-stellar disks in young star clusters, Thies et al. (2010) find that a theoretical BD and VLMS MF emerges which is the same as the observationally deduced one.

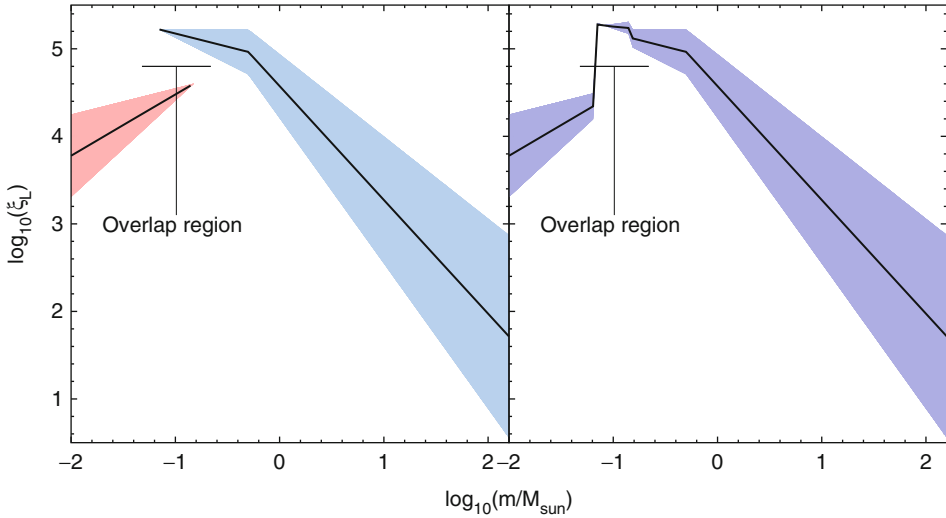


Fig. 4-23

Left frame: The canonical IMF (solid line, (4.55)) with separate BD and stellar components. The upper and lower envelopes of the shaded regions are the minimum and maximum IMF when the uncertainties are taken into account. Both components overlap between 0.07 and $0.15 M_{\odot}$. *Right frame:* The sum of the BD and the stellar components as it would appear to an observer if perfect mass-measurements were available of all stars and BDs in all multiple systems. Note the “bump” that marks the overlap region. However, as demonstrated in Fig. 4-22, the real observational data wash out such features owing to the measurement uncertainties and the errors in transforming observed fluxes to masses. The model mass-histograms (thick red lines in Fig. 4-22) then fit the observed mass-histograms excellently for the young clusters for which such data are available

The observational and theoretical result that the BD branch becomes insignificant (i.e., the BD MF falls off steeply) above $0.1\text{--}0.2 M_{\odot}$ follows from a combination of two effects: the limited amount of material around late-type pre-main sequence stars that can be accreted onto a gravitational instability within the outer region of a circum-protostellar mass and the rare occurrence of massive protostars which are likely to have massive disks.

Main Results

Brown dwarfs are a separate population when compared to stars due to their distinct binary properties. The BD IMF has a power-law index $\alpha_0 \approx 0.3$, and the IMF is discontinuous but with a significant overlap of masses between the sub-stellar and stellar regime. This may hide the discontinuity unless a proper analysis is done. About one BD forms per five stars.

9 The Shape of the IMF from Resolved Stellar Populations

From the above discourse, it thus becomes apparent that we have good constraints on the stellar and BD IMF. These are valid only for the regime of present-day (“normal”) star formation, that is, star-formation densities $\rho \lesssim 10^5 M_{\odot} \text{pc}^{-3}$ and metallicities $[\text{Fe}/\text{H}] \gtrsim -2$. Since stars are formed

as binary systems (➤ Sect. 2.6), the system IMF is provided in ➤ Sect. 9.2 for binary fractions of 100% (the birth system IMF) and for a binary fraction of 50% (the typical Galactic-field or open star cluster system IMF). The Galactic-field IMF, which is the IGIMF valid for the Milky Way, is provided in ➤ Sect. 9.3. In ➤ Sect. 12, it is concluded that the IMF becomes top-heavy for star formation under denser conditions and that it may be bottom-light under metal-poor conditions.

The distribution of stars by mass in “normal” systems is a power law with exponent or index $\alpha_2 \approx 2.3$ for stellar masses $m \gtrsim 0.5 M_\odot$. There exists a physical stellar mass limit, $m_{\text{max}^*} \approx 150 M_\odot$ such that $m \leq m_{\text{max}^*}$ (➤ Sect. 3). The distribution of stars below the K/M dwarf transition mass, $0.5 M_\odot$, can also be described by a power law but with exponent $\alpha_1 \approx 1.3$ (➤ Sect. 7). Given the latest results described in ➤ Sect. 8, the mass distribution below the mass $m_1 \approx 0.1 M_\odot$ is uncertain, but measurements indicate a power law with exponent $0 < \alpha_0 < 1$. Because the binary properties of VLMSs and BDs differ substantially from those in the low-mass star regime, it emerges stringently that BDs and some VLMSs need to be considered as a separate population that is linked to, but different from stars. Fitting a functional description of the mass distribution with the continuity constraint across m_1 would therefore be wrong. It follows that one single function such as the log-normal form, which may be associated with the likelihood of occurrence of masses from the fragmentation limit,¹⁵ $m_0 \approx 0.01 M_\odot$, through to the physical stability limit, m_{max^*} , is not the correct description of the stellar and BD IMF.

With these recent insights (power-law IMF over two orders of magnitude in mass and discontinuity near the sub-stellar mass limit), little of the argument for the advantages of a log-normal or any other mathematical form (➤ Table 4-3 below) remains. Indeed, any such other mathematical form has the disadvantage that the tails of the distribution react to changes in the parametrization in a way perhaps not wanted when testing models. To give an example, a single log-normal form would change the slope of the IMF at large masses even if only the LF for late-type stars is to be varied. The canonical (➤ 4.55) two-part power-law stellar IMF, on the other hand, would allow changes to the index at low masses without affecting the high-mass end, and the addition of further power-law segments is mathematically convenient. The canonical two-part power-law stellar IMF also captures the essence of the physics of star formation, namely, a featureless power-law form for the largest range of stellar masses and a turnover near some fraction of a solar mass. This turnover appears to be present already in the pre-stellar cloud-core MF and may be due to the decreasing likelihood that low-mass cloud clumps collapse under self-gravity (➤ Sect. 11.2).

9.1 The Canonical, Standard or Average IMF

The various constraints arrived at above are summarized by an IMF that is a single power law for BDs and a two-part power law for stars (➤ 4.55), using the notation from (➤ 4.4) to (➤ 4.5).

¹⁵When a cloud collapses, its density increases, but its temperature remains constant as long as the opacity remains low enough to enable the contraction work to be radiated away. The Jeans mass (➤ 4.50) consequently decreases and further fragments with smaller masses form. When, however, the density increases to a level such that the cloud core becomes optically thick, then the temperature increases, and the Jeans mass follows suit. Thus, an opacity-limited minimum fragmentation mass of about $0.01 M_\odot$ is arrived at (Low and Lynden-Bell 1976; Boss 1986; Kumar 2003; Bate 2005).

The Canonical IMF $(m$ is in units of M_\odot)

$$\begin{aligned} \xi_{\text{BD}}(m) &= \frac{k}{3} \left(\frac{m}{0.07}\right)^{-0.3 \pm 0.4} & , \quad 0.01 < m \lesssim 0.15, \\ \xi_{\text{star}}(m) &= k \begin{cases} \left(\frac{m}{0.07}\right)^{-1.3 \pm 0.3} & , \quad 0.07 < m \leq 0.5, \\ \left[\left(\frac{0.5}{0.07}\right)^{-1.3 \pm 0.3}\right] \left(\frac{m}{0.5}\right)^{-2.3 \pm 0.36} & , \quad 0.5 < m \leq 150. \end{cases} \end{aligned} \quad (4.55)$$

The uncertainties are discussed in [Sect. 9.4](#). This is the individual-star/BD IMF which is corrected fully for multiple companions. That is, in a star-formation event, the distribution of stars and of BDs that form is given by this IMF. The constraint that the population be optimally sampled (p. 132) may be invoked. To formulate a realistic stellar and BD population would then require these stars and BDs to be distributed into stellar and BD binaries ([Sects. 2.6](#) and [9.2](#)).

Note that this form is a two-part power law in the stellar regime and that BDs contribute about 4% by mass only and need to be treated as a separate population such that both IMFs overlap between about 0.07 and 0.15 M_\odot ([Fig. 4-23](#)). The gap or discontinuity between the BD and the stellar IMF can be measured by the BD-to-star ratio at the classical BD-star border, $m_{\text{H}} \approx 0.075 M_\odot$, with $\xi_{\text{BD}}(m_{\text{H}})/\xi_{\text{star}}(m_{\text{H}}) \approx \frac{1}{3}$, that is, $k_{\text{BD}} \approx 1/3$ (Kroupa et al. 2003; Thies and Kroupa 2007, 2008).

The canonical IMF can also be described as a log-normal function for low-mass stars with a power-law extension to massive stars, yielding the *log-normal canonical IMF*.

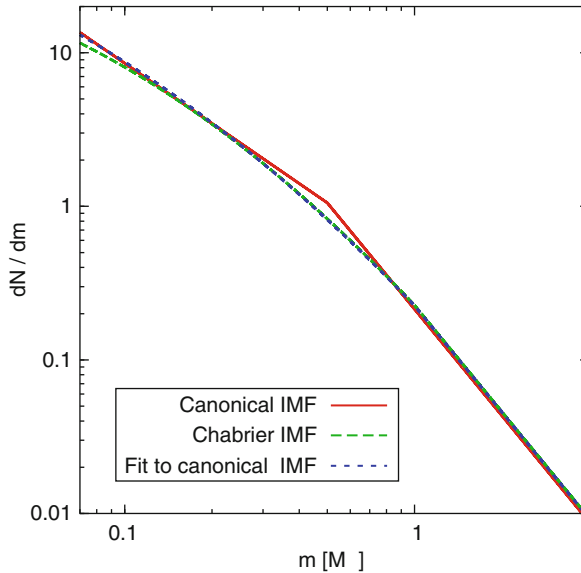
The Log-normal Canonical IMF $(m$ is in units of M_\odot)

$$\begin{aligned} \xi_{\text{BD}}(m) &= k k_{\text{BD}} \left(\frac{m}{0.07}\right)^{-0.3 \pm 0.4} & , \quad 0.01 < m \lesssim 0.15, \\ \xi_{\text{star}}(m) &= k \begin{cases} \frac{1}{m} \exp\left[-\frac{(lm - lm_c)^2}{2\sigma_{lm}^2}\right] & , \quad 0.07 < m \leq 1.0, \\ A \left(\frac{m}{1.0}\right)^{-2.3 \pm 0.36} & , \quad 1.0 < m \leq 150. \end{cases} \end{aligned} \quad (4.56)$$

In [\(4.56\)](#) $lm_c \equiv \log_{10} m_c / M_\odot$, continuity is assured at 1 M_\odot and $\xi_{\text{BD}}(m_{\text{H}})/\xi_{\text{star}}(m_{\text{H}}) \approx \frac{1}{3}$, as in [\(4.55\)](#).

A least-squares fit of the log-normal canonical IMF to the two-part power-law form ([4.55](#)), whereby $\int_{0.07}^{150} m \xi(m) dm = 1 M_\odot$ for both (m in Solar units), yields $m_c = 0.055 M_\odot$ and $\sigma_{lm} = 0.75$ with $A = 0.2440$ for continuity at 1 M_\odot and $k_{\text{BD}} = 4.46$ to ensure $\xi_{\text{BD}}(0.75 M_\odot)/\xi_{\text{star}}(0.75 M_\odot) = 1/3$ as being the best log-normal plus power-law representation of the canonical IMF. Alternatively, the Chabrier parametrization has $m_c = 0.079_{+0.021}^{-0.016} M_\odot$ and $\sigma_{lm} = 0.69_{+0.05}^{-0.01}$ (Table 1 in Chabrier 2003a) with $A = 0.2791$ and $k_{\text{BD}} = 4.53$.

The three forms of the canonical IMF are compared in [Fig. 4-24](#). The best-fit log-normal representation of the two-part power-law canonical IMF is indistinguishable from the Chabrier result, demonstrating the extreme robustness of the canonical IMF.



■ Fig. 4-24

A comparison between the three canonical IMFs: the two-part power-law IMF (([4.55](#)), *solid red curve*) and the log-normal plus power-law IMF (([4.56](#)), the “best-fit-canonical IMF” and the “Chabrier IMF”) in the interval from 0.07 to $4 M_{\odot}$. Plotted is the number of stars per mass interval versus the stellar mass, both scales being logarithmic. The IMFs are normalized such that $\int_{0.07}^{150} m \xi(m) dm = 1 M_{\odot}$, where m is the mass in solar units. Note in particular that the three IMF forms are indistinguishable over the whole mass interval. They are identical above a mass of $1 M_{\odot}$, except for a slightly different normalization factor

The above canonical or standard forms have been derived from detailed considerations of star counts thereby representing an *average* IMF: for low-mass stars, it is a mixture of stellar populations spanning a large range of ages (0–10 Gyr) and metallicities ($[\text{Fe}/\text{H}] \gtrsim -2$). For the massive stars, it constitutes a mixture of different metallicities ($[\text{Fe}/\text{H}] \gtrsim -1.5$) and star-forming conditions (OB associations to very dense star-burst clusters: R136 in the LMC). Therefore, the average IMF can be taken as a canonical form, and the aim is to test the IMF Universality Hypothesis:

IMF Universality Hypothesis

The canonical IMF ([4.55](#)) and ([4.56](#)) constitutes the parent distribution of all stellar populations that form with densities $\rho \lesssim 10^5 M_{\odot}/\text{pc}^3$ and metallicities $[\text{Fe}/\text{H}] \gtrsim -2$.

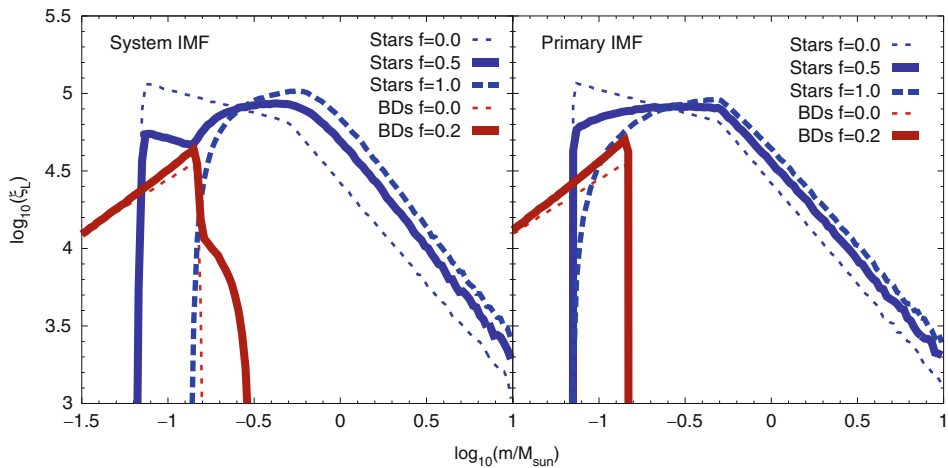
Negation of this hypothesis would indeed imply a variable IMF. For larger densities, evidence has emerged that the IMF becomes top-heavy ([4.64](#)). Also, evidence has emerged that the IMF may have a metallicity dependence ([4.63](#)).

9.2 The IMF of Systems and of Primaries

The canonical stellar IMF (☉ 4.55) is the distribution of all stars to have formed together in one star-forming event. However, since stars form as binary systems, it is also useful to consider the form of the system IMF which results from pairing stars chosen randomly from the canonical stellar IMF. Random pairing is a good description of late-type stellar systems in birth environments, but massive stars tend to prefer more similar-mass companions (☉ Sect. 2.6). The system IMF is given by (☉ 4.57) by approximating the mass distribution by a power law with index α_1 between 0.15 and $0.5 M_\odot$. It is assumed that all stars are in binaries, that is, $f = 1$, and, for completeness, the system IMF for a binary fraction of 50% is also evaluated. The actual mass distributions are plotted in ☉ Fig. 4-25.

The system IMF:

$$\begin{aligned} \alpha_1 &= +0.66 \pm 0.3 & , & \quad 0.15 \leq m/M_\odot < 0.65 & , & \quad (f = 0.5) \\ \alpha_1 &\approx -0.22 & , & \quad 0.15 \leq m/M_\odot < 0.65 & , & \quad (f = 1). \end{aligned} \quad (4.57)$$



☐ Fig. 4-25

Left panel: The system IMF is shown in blue and the system BD/VLMS IMF is in red. The *thick curves* are for a binary fraction among stars of $f = 50\%$ and for a BD binary fraction of $f_{\text{BD}} = 20\%$. Random pairing of companion masses is assumed (☉ Sect. 2.6) from the canonical stellar ($0.07\text{--}150 M_\odot$) and BD ($0.01\text{--}0.15 M_\odot$) IMFs. These are shown here as the *thin dotted curves* for binary fraction = 0 ((☉ 4.55), also plotted in ☉ Fig. 4-23). Note the flat extension at the lowest-mass end of the stellar system IMF. It comes from the fact that below $0.14 M_\odot$, all stars are single since the minimum system mass is twice the minimum stellar mass ($0.07 M_\odot$). The bump at the high-mass end of the BD system distribution comes from the most massive systems having masses larger than the most massive BD/VLMS ($0.15 M_\odot$) in this model. The medium dashed stellar system IMFs are for an initial stellar binary fraction of 100%, as would be found in dynamically not evolved star-forming regions. **Right panel:** the same but plotting only the IMF of primary masses. Note the significant difference between the canonical IMF and the system/primary-star IMFs in both panels. The system IMF and the IMF of primaries are approximated by (☉ 4.57) and (☉ 4.58), respectively

Note that in each case $\alpha_0 = 0.2 \pm 0.4$ ($f_{\text{BD}} = 0.2$, random pairing) and $\alpha_2 = 2.3$ ($m \gtrsim 0.65 M_\odot$) as in the canonical stellar IMF (● 4.55).

Note that the binary effect contributes $\Delta\alpha \approx 0.64$ for $f = 0.5$ and $\Delta\alpha \approx 1.5$ for $f = 1$ where $\Delta\alpha$ is the difference between the canonical $\alpha_1 = 1.3$ and the above values.

The IMF of primary stars, which may be closer to an observed IMF since companions are usually faint and would be lost from the star count, is given by (● 4.58) following the same procedure as for the system IMF.

The IMF of primary stars:

$$\begin{aligned} \alpha_1 &= +1.0 \pm 0.3 & , & \quad 0.15 \leq m/M_\odot < 0.50 & , & \quad (f = 0.5) \\ \alpha_1 &= +0.7 \pm 0.3 & , & \quad 0.15 \leq m/M_\odot < 0.50 & , & \quad (f = 1). \end{aligned} \quad (4.58)$$

Note that in each case $\alpha_0 = 0.2 \pm 0.4$ ($f_{\text{BD}} = 0.2$, random pairing) and $\alpha_2 = 2.3$ ($m \gtrsim 0.5 M_\odot$) as in the canonical stellar IMF (● 4.55).

9.3 The Galactic-Field IMF

The Scalo power-law index $\alpha_3 = 2.7$ ($m \gtrsim 1 M_\odot$) was inferred by Scalo (1986) from star counts for the MW disk population of massive stars such that the three-part power-law KTG93 IMF needs to be identified with the *composite IMF* (i.e., the IGIMF) introduced in ● Sect. 13.1, arriving at the *KTG93 IMF* (Kroupa et al. 1993). It is updated in (● 4.59) to account for the separate BD and stellar populations.

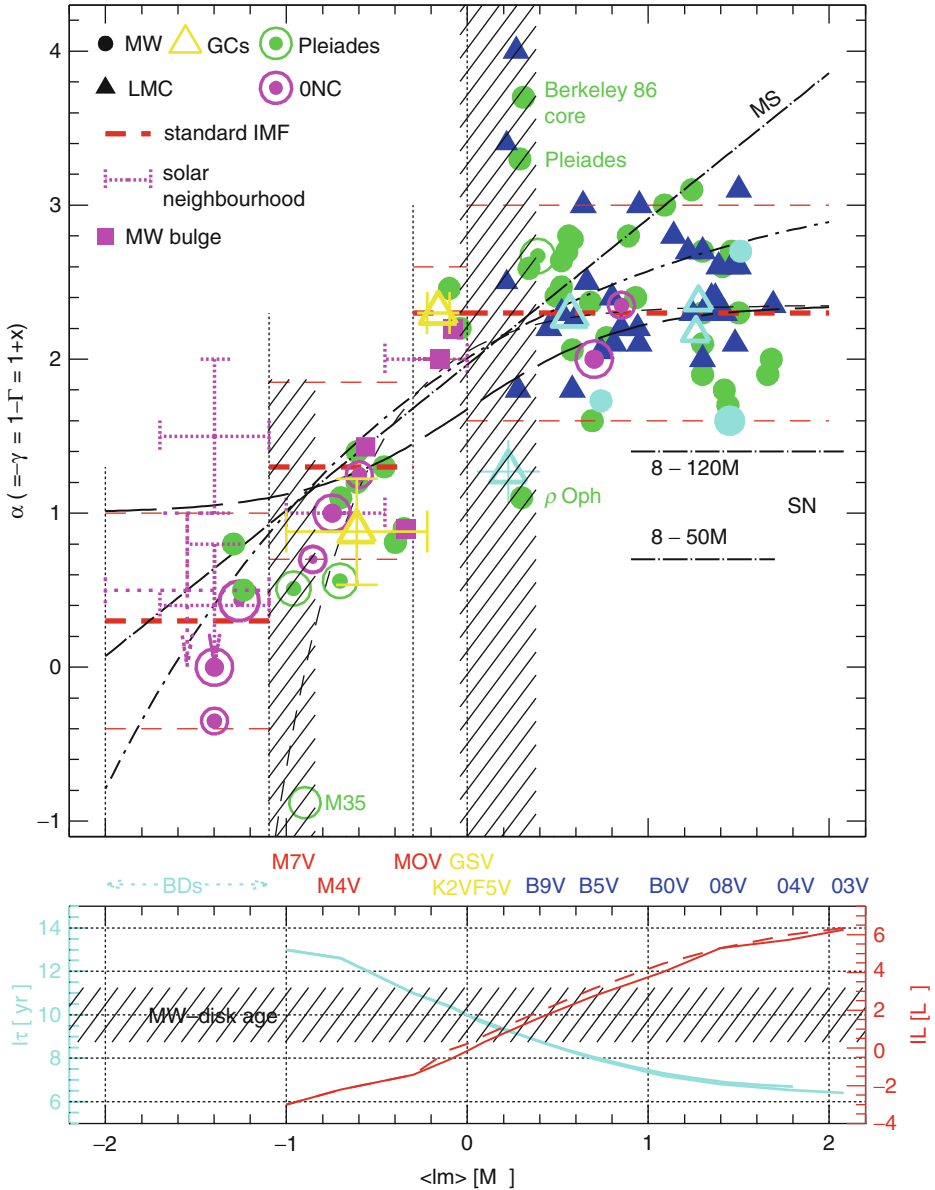
The Galactic-field (KTG93) stellar IMF:

$$\begin{aligned} \alpha_0 &= +0.3 \pm 0.4 & , & \quad 0.01 \leq m/M_\odot \lesssim 0.15 & , & \quad n = 0, \text{ (BDs)} \\ \alpha_1 &= +1.3 \pm 0.3 & , & \quad 0.07 \lesssim m/M_\odot < 0.50 & , & \quad n = 1, \text{ (stars)} \\ \alpha_2 &= +2.3 \pm 0.3 & , & \quad 0.5 \leq m/M_\odot < 1 & , & \quad n = 2, \\ \alpha_3 &= +2.7 \pm 0.4 & , & \quad 1 \lesssim m/M_\odot & , & \quad n = 3. \end{aligned} \quad (4.59)$$

The Galactic-field (KTG93) IMF is the IMF of all young stars within the Galactic disk assuming all multiple stars can be resolved. It is not equal to the canonical IMF for $m \gtrsim 1 M_\odot$. In the LMC, the field-star IMF is known to be steep, $\alpha_3 \approx 4.5$ (Massey 2003).

9.4 The Alpha Plot

A convenient way for summarizing various studies of the IMF is by plotting independently derived power-law indices in dependence of the stellar mass over which they are fitted (Scalo 1998; Kroupa 2001b, 2002; Hillenbrand 2004). The upper panel of ● Fig. 4-26 shows such data: The shape of the IMF is mapped by plotting measurements of α at $\langle lm \rangle = (lm_2 - lm_1)/2$ obtained by fitting power laws, $\xi(m) \propto m^{-\alpha}$, to logarithmic mass ranges lm_1 to lm_2 (not indicated here for clarity). Circles and triangles are data compiled by Scalo (1998) and Kroupa (2001b) for MW and Large-Magellanic-Cloud (LMC) clusters and OB associations, as well as newer data, some of which are emphasized using different symbols (and colors). Unresolved multiple systems are not corrected for in all these data including the MW-Bulge data. The canonical stellar IMF (● 4.55), corrected for unseen binary-star companions, is the two-part power law (thick short-dashed lines). Other binary-star-corrected solar-neighborhood IMF measurements are indicated as (magenta) dotted error bars.



■ Fig. 4-26

Upper panel: the alpha plot. The curves (e.g., labeled “MS”) are various IMF slopes documented in [Table 4-3](#) and discussed in [Sect. 10.2](#). The plotted data are listed in the supplementary information of [Kroupa \(2002\)](#). The cyan open triangles are for R136 in the LMC. **Lower panel:** The bolometric stellar mass–luminosity relation, $IL(lm)$, shown by the *solid* and *dashed* curves, and the stellar main-sequence lifetime (or turnoff masses at a given age), $l\tau$. The possible range of Milky Way (MW) disk ages are shown as the *shaded region*. The masses of stellar spectral types are indicated. Notation: $lm \equiv \log_{10}(m/M_{\odot})$, $l\tau \equiv \log_{10}(\tau/\text{year})$, $IL \equiv \log_{10}(L/L_{\odot})$. For more details, see text

In the lower panel of [Fig. 4-26](#) are plotted the luminosities of main-sequence stars and the stellar lifetimes as a function of mass on the same mass scale as the alpha plot shown in the upper panel.

For $m > 1 M_{\odot}$, correction for unseen companions does not affect the IMF (Maíz Apellániz 2008; Weidner et al. 2009). The M dwarf ($0.1\text{--}0.5 M_{\odot}$) MFs for the various clusters are systematically flatter (smaller α_1) than the canonical IMF, which is mostly due to unresolved multiple systems in the observed values. This is verified by comparing to the system IMF for a binary fraction of 50% ([Fig. 4.57](#)). Some of the data do coincide with the canonical IMF though, and Kroupa (2001b) argues that on correcting these for unresolved binaries, the underlying true individual-star IMF ought to have $\alpha_1 \approx 1.8$. This may indicate a systematic variation of α_1 with metallicity because the data are young clusters that are typically more metal-rich than the average Galactic field population for which $\alpha_1 = 1.3$ ([Fig. 4.62](#) below).

A power-law extension into the BD regime with a smaller index ($\alpha_0 = +0.3$) is shown as a third thick short-dashed segment, but this part of the mass distribution is not a continuous extension of the stellar distribution, as noted in [Sect. 8.4](#).

The upper and lower thin short-dashed lines are the estimated 99% confidence range on α_i . The usual one-sigma uncertainties adopted in [Fig. 4.55](#) are, however, estimated from the distribution of α values in [Figs. 4-26](#) and [Fig. 4-27](#).

The long-dash-dotted horizontal lines in [Fig. 4-26](#) labeled “SN” are those IMFs with $\alpha_3 = 0.70(1.4)$ but $\alpha_0, \alpha_1, \alpha_2$ as in [Fig. 4.55](#), for which 50% of the stellar (including BD) mass is in stars with 8–50 (8–120) M_{\odot} , respectively. It is noteworthy that none of the available resolved clusters, not even including the Local Group star-burst clusters, have such a top-heavy IMF.

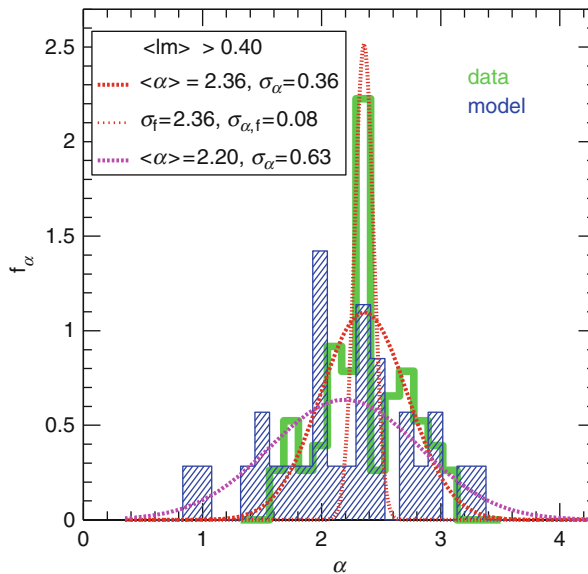


Fig. 4-27

Distribution of α values for massive stars ($m > 10^{0.4} M_{\odot}$). Note that the empirical data show a sharper (just at the Salpeter index) and more symmetrical distribution than the model, which is based on the Salpeter index. This is unexpected. See Open Question III in [Sect. 9.5](#) for details

The vertical dotted lines in [Fig. 4-26](#) delineate the four mass ranges over which the Galactic-field IMF is defined ([4.59](#)), and the shaded areas highlight those stellar mass regions where the derivation of the IMF is additionally complicated especially for Galactic field stars: For $0.07 < m/M_{\odot} < 0.15$, long pre-main sequence contraction times (Chabrier and Baraffe 2000) make the conversion from an empirical LF to an IMF ([4.1](#)) dependent on the precise knowledge of stellar ages and the star-formation history, and for $0.8 < m/M_{\odot} < 2.5$ uncertain main-sequence evolution, Galactic-disk age and the star-formation history of the MW disk do not allow accurate IMF determinations (Binney et al. 2000; Elmegreen and Scalo 2006).

9.5 The Distribution of Data Points in the Alpha-Plot

The first thing to note about the data distribution in the alpha plot is that there is no readily discernible systematic difference in IMF determinations neither with metallicity nor density of the population (cf. [Fig. 4-8](#)).

In order to understand the origin and nature of the dispersion of power-law indices evident in the alpha plot, Kroupa (2001b) investigates the dispersion of α values for a given mass interval using Aarseth- N -body models of evolving clusters. The result is that the dispersion can be understood in terms of statistical sampling from a universal IMF (as also found by Elmegreen 1997, 1999) together with stellar-dynamical biases.¹⁶ Optimal Sampling ([Sect. 2.2](#)) was not available in 2001 so that statistical variations that arise by randomly sampling stars from the IMF could not then be separated from variations of the IMF due to stellar-dynamical processes in young star clusters.

Given the existing 2001 theoretical investigation, it is possible to compare the theoretical distribution of α values for an ensemble of star clusters with the observational data. This is done for stars with $m > 2.5 M_{\odot}$ in [Fig. 4-27](#) where the open (green) histogram shows the distribution of observational data from [Fig. 4-26](#). The (blue) shaded histogram is the theoretical ensemble of 12 star clusters containing initially 800 to 10^4 stars that are “observed” at 3 and 70 Myr: Stellar companions are merged to give the system MFs, which are used to measure α , but the input individual-star IMF is in all cases the canonical form ([4.55](#)). The dotted curves are Gaussians with the same average α and standard deviation in α obtained from the histograms. Fixing $\alpha_f = \langle \alpha \rangle$ and using only $|\alpha| \leq 2\sigma_{\alpha}$ for the observational data gives the narrow thin (red) dotted Gaussian distribution which describes the Salpeter peak extremely well (not by construction).

The interesting finding is thus that the observational data have a very pronounced Salpeter/Massey peak, with broad near-symmetric wings. This indicates that there are no significant biases that should skew the distribution. For example, if the observational sample contained clusters or associations within which the OB stars have a low binary fraction compared to others that have a very high multiplicity fraction, we would expect the binary-deficient cases to deviate toward high α values since fainter companions are hidden in the binary-rich cases. The absence of this effect is consistent with the result obtained by Maíz Apellániz (2008) and Weidner et al. (2009) that multiple systems do not affect the derived power-law index of the IMF for stars more massive than a few M_{\odot} . For stars with mass $m < 1 M_{\odot}$, unresolved multiples do have a significant effect, and this has been corrected for in the canonical IMF ([4.55](#)).

¹⁶Note that this does *not* constitute a proof of the stellar IMF being a probabilistic density distribution!

Energetic dynamical ejections from cluster cores deplete the IMF of the massive stars (Pflamm-Altenburg and Kroupa 2006), increasing the observed α_3 , (Banerjee and Kroupa 2012) in the cluster while mass segregation has the opposite effect.

In contrast to the observational data, the theoretical data show (i) a distribution with a mean shifted to smaller $\alpha_3 \approx 2.2$ that has (ii) a larger width than the observational one. The input canonical Salpeter/Massey index is not really evident in the theoretical data, and if these were the observational data, then it is likely that the astronomical community would strongly argue for the case that the IMF shows appreciable variations. This leads to the following peculiar result.

Open Question III

The empirical IMF power-law indices for massive stars are better behaved than the model data. This is unexpected because all the additional complications (observational uncertainties, uncertainties in transferring observed luminosities to masses such as stellar models, rotating/non-rotating stars) ought to deteriorate the empirical data, while the model has none of these uncertainties.

Clarifying the hitherto not understood difference between the much more “well-behaved” observational data and the theoretical data will need further theoretical work which will have to attempt to reproduce the observational procedure as exactly as is possible (see also the Sociological Hypothesis on p. 196).

Note that Open Question III is naturally resolved if nature follows optimal sampling (☛ Sect. 2.2) rather than random sampling (☛ Sect. 2.3) from the IMF.

Is the scatter of data points in the alpha plot (☛ Fig. 4-26) a result of IMF variations? For this to be conclusively convincing would require a measurement to lie further from the canonical IMF than the conservative uncertainty range shown in the figure. However, the adopted uncertainties on α_i in (☛ 4.55) stem from the scatter in the alpha plot, so that this argument is circular.

An independent indication of the uncertainties inherent to IMF determinations can be obtained by comparing IMF estimates of the same cluster by different authors. This is demonstrated for the well-studied Pleiades, ONC, and for 30 Dor (☛ Fig. 4-26). Overall, the uncertainties in α are about ± 0.5 which is also about the scatter evident in all the data, so that there is no indication of significant outliers (except in the shaded regions, see below). Differences of $\Delta\alpha \approx 0.5$ for VLMSs and BDs are evident for the extremely young ONC allowing an estimate of likely nonphysical variations in the alpha plot. Data reduction at these low masses is hampered by variability, differential reddening, and spurious detections (Slesnick et al. 2004; Andersen et al. 2009). It is clear that because the procedure of measuring $\alpha(lm)$ is not standardized and because the IMF is not a single power law, author–author variations occur simply due to the use of different mass ranges when fitting power laws (see “Binning” bias in ☛ Sect. 2.1).

Significant departures from the canonical IMF only occur in the shaded areas of the alpha plot. These are, however, not reliable. The upper mass range in the shaded area near $1 M_{\odot}$ poses the problem that the star clusters investigated have evolved such that the turnoff mass is near to this range. Some clusters such as ρ Oph are so sparse that more massive stars did not form. In these cases, the shaded range is close to the upper mass limit leading to possible stochastic stellar-dynamical biases since the most massive stars meet near the core of a cluster due to mass segregation, but three-body or higher-order encounters there can cause expulsion from

the cluster. Furthermore, ρ Oph is still forming, leading to unknown effects that are likely to enhance variations in the first derivative of the IMF (i.e., in α values).

The shaded area near $0.1 M_{\odot}$ poses the problem that the VLMSs are not on the main sequence for most of the clusters studied and are again prone to bias through mass segregation by being underrepresented within the central cluster area that is easiest to study observationally. Especially the latter is probably biasing the M35 datum, but some effect with metallicity may be operating especially so since M35 appears to have a smaller α near the H-burning mass limit than the Pleiades cluster which has a similar age but has a larger abundance of metals (► Fig. 4-19). The M35 cluster ought to be looked at again with telescopes.

Two other well-studied massive star-burst clusters have $\alpha \approx \text{Salpeter/Massey}$ (30 Dor and NGC 3603) implying no clear evidence for a bias that resolved star-burst clusters prefer smaller α and thus more massive stars relatively to the number of low-mass stars. Low-mass stars are known to form in 30 Dor (Sirianni et al. 2000), although their MF has not been measured yet due to the large distance of about 55 kpc. From the ONC, we know that the entire mass spectrum $0.05 \lesssim m/M_{\odot} \lesssim 40$ is represented but that it has a deficit of massive stars (Pflamm-Altenburg and Kroupa 2006) (► Fig. 4-19). The Pleiades appear to have had an IMF very similar to the canonical one, although for massive stars, a steeper IMF with $\alpha_3 \approx 2.7$ may also be suggested by theoretical work (► Fig. 4-18).

But it remains an unsolved issue (see Open Question III on p. 195) as to why the theoretical data have a larger dispersion of α values than the empirical ones (► Fig. 4-27). The models discussed above and plotted in ► Fig. 4-27 had as the input IMF the Salpeter index, whereas the empirical data suffer under all the biases associated with transforming measured fluxes to stellar masses (► Sect. 2.1) and have been collated from various sources published by different authors. In particular, the very narrow empirical distribution exactly around the Salpeter index is remarkable. This leads to the following Sociological Hypothesis.

Sociological Hypothesis

The measured α indices for massive stars are affected by sociological predispositions.

However, the excellent documented efforts of the teams working on measuring α for massive stars would not necessarily lend support to this hypothesis. Actually, as already observed on p. 195, optimal sampling would automatically resolve Open Question III therewith negating the Sociological Hypothesis.

The available evidence is thus that low-mass stars and massive stars form together even in extreme environments. This is also supported by an impressive observational study (Luhman et al. 2000; Luhman 2004) of many close-by star-forming regions using one consistent methodology to avoid author–author variations. The result is that the IMF does not show conclusive differences from low-density star-forming regions in small molecular clouds ($n = 0.2\text{--}1$ stars/pc³ in ρ Oph) to high-density cases in giant molecular clouds ($n = (1 - 5) \times 10^4$ stars/pc³ in the ONC). This result extends to the populations in the truly exotic ancient and metal-poor dwarf-spheroidal satellite galaxies which are speculated to be dominated by dark matter but definitely constitute star-forming conditions very different from present-day events. Two such close companions to the Milky Way have been observed (Grillmair et al. 1998; Feltzing et al. 1999) finding the same MF as in globular clusters for $0.5 \lesssim m/M_{\odot} \lesssim 0.9$ and thus no evident differences

to the canonical IMF. However, evidence for top-heavy IMFs in pc-scale starbursts and for bottom-heavy IMFs in elliptical galaxies has emerged (► Sect. 12.9 and ► 12.3, respectively).

Main Results

Within the observational uncertainties, the IMF of all known resolved stellar populations is well described by the canonical, standard, or average IMF. It is given by (► 4.55) and has been corrected for unresolved multiple systems. The only structure evident in the *stellar* IMF is thus a turnover near $0.5 M_{\odot}$ and a rapid turndown below $0.075 M_{\odot}$. The BD IMF is a separate single power-law such that about one BD forms per five stars. Open Question III leads to the Sociological Hypothesis which may naturally be negated by optimal sampling.

10 Comparisons and Some Numbers

In this section, some useful numbers are provided, and a comparison between various IMF forms is made with cumulative functions in the number of stars and stellar mass being plotted for general use.

10.1 The Solar-Neighborhood Mass Density and Some Other Numbers

Given the reasonably well-constrained shape of the stellar IMF (► 4.55), it is of interest to investigate the implied number and mass density in the Galactic disk. Here, we consider the solar neighborhood. In order to normalize the IMF to the solar neighborhood stellar number density, the observed stellar LF is conveniently used.

The nearby Hipparcos LF, $\Psi_{\text{near}}(\text{Hipp})$ (► Fig. 4-9), has $\rho = (5.9 \pm 0.3) \times 10^{-3}$ stars/pc³ in the interval $M_V = 5.5 - 7.5$ corresponding to the mass interval $[m_2, m_1] = [0.891, 0.687] M_{\odot}$ (Kroupa 2001c) using the KTG93 MLR (► Fig. 4-11). $\int_{m_1}^{m_2} \xi(m) dm = \rho$ yields $k = 0.877 \pm 0.045$ stars/(pc³ M_{\odot}). The number fractions, mass fractions, and Galactic-field mass densities contributed by stars in different mass ranges are summarized in ► Table 4-1 (p. 161).

The local mass density made up of interstellar matter is $\rho^{\text{gas}} \approx 0.04 \pm 0.02 M_{\odot}/\text{pc}^3$. In stellar remnants, it is $\rho^{\text{rem}} \approx 0.003 M_{\odot}/\text{pc}^3$ (Weidemann 1990) or $\rho^{\text{rem}} \approx 0.005 M_{\odot}/\text{pc}^3$ (Chabrier 2003a and references therein). Giant stars contribute about $0.6 \times 10^{-3} M_{\odot}/\text{pc}^3$ (Haywood et al. 1997), so that main-sequence stars make up about half of the baryonic matter density in the local Galactic disk (► Table 4-1). BDs, which for some time were regarded as candidates for contributing to the dark-matter problem, do not constitute a dynamically important mass component of the Galaxy, contributing not more than 5% in mass. This is corroborated by dynamical analysis of local stellar space motions that imply there is no need for dark matter in the Milky Way disk (Flynn and Fuchs 1994), and the revision of the thick-disk mass density to larger values (Fuhrmann 2004; Soubiran et al. 2003) further reduces the need for dark matter within the solar circle.

► *Table 4-1* also shows that a star cluster loses about 12% of its mass through stellar evolution within 10 Myr if $\alpha_3 = 2.3$ (turnoff mass $m_{t0} \approx 20 M_\odot$) or within 300 Myr if $\alpha_3 = 2.7$ (turnoff-mass $m_{t0} \approx 3 M_\odot$). After 5 Gyr, the mass loss through stellar evolution alone amounts to about 45% if $\alpha_3 = 2.3$ or 30% if $\alpha_3 = 2.7$. Mass loss through stellar evolution therefore poses no risk for the survival of star clusters for the IMFs discussed here since the mass-loss rate is slow enough for the cluster to adjust adiabatically. A star cluster would be threatened through mass loss from supernova explosions if $\alpha \lesssim 1.4$ for $8 < m/M_\odot \leq 120$ which would mean a mass loss of $\gtrsim 50\%$ within about 40 Myr when the last supernova explodes. It is remarkable that none of the measurements in resolved populations has found such a low α for massive stars (► *Fig. 4-26*). Mass loss due to stellar evolution might pose a threat to the survival of post-gas-expulsion clusters when the system is mass-segregated as has been shown by Vesperini et al. (2009).

10.2 Other IMF Forms and Cumulative Functions

The standard or canonical power-law IMF (► 4.55) provides a good description of the data combined with mathematical ease and physical meaning. Integrating the IMF is a frequently required task, and it is especially here that the two- or even multi-part power-law description of the canonical IMF shows its strength. For example, if the IMF were a probabilistic density distribution function then writing down a mass-generating function (► Sect. 2.3) is practically trivial allowing perfectly efficient (each random deviate X yields a usable mass m) discretization of the stellar population into individual stellar masses. A further strong advantage of this parametrization is that each section can be changed without affecting another part of the IMF. As an explicit example, should the BD MF be revised, α_0 can be adopted accordingly without affecting the rest of the mass distribution in the well-constrained stellar regime. That the two- or multi-part power-law form has a discontinuity in the derivative at $0.5 M_\odot$ (► *Fig. 4-26*) has no implications for stellar populations and is therefore not a disadvantage, especially so since there exists no theoretically derived IMF form of significance. It is, however, true that the real IMF must be differentiable, but it must have a rapid change in slope near $0.5 M_\odot$ and below about $0.08 M_\odot$ where the stellar IMF decays rapidly. The Chabrier formulation of the log-normal canonical IMF ((► 4.56), ► *Fig. 4-24*) also has a discontinuity in its derivative but at $1 M_\odot$ and does not reproduce the rapid falloff of the stellar IMF below about $0.08 M_\odot$ unless it is cutoff there nor is it easily integrable.

Often, a single Salpeter power law IMF is applied ($\alpha = 2.35$ for $0.1 M_\odot \lesssim m$). ► *Table 4-2* gives the stellar masses (► 4.8) for the single power-law Salpeter IMF in comparison to the canonical IMF. For example, using a single power-law “Salpeter” IMF with $\alpha = 2.3$ for a stellar population with stars in the mass range $0.1\text{--}0.8 M_\odot$ instead of the canonical IMF would lead to an overestimate of the stellar mass by $0.66/0.37 = 78\%$ and to an overestimate of the number, N , of stars by a factor $1.687/0.702 = 2.4$.

Additional forms are in use and are preferred for some investigations: In ► *Fig. 4-26*, the quasi-diagonal (black) lines are alternative analytical forms summarized in ► *Table 4-3*. They are compared to the canonical IMF in ► *Fig. 4-28*, and their cumulative number and cumulative mass functions are presented in ► *Figs. 4-29* and ► 4-30, respectively.

Of the other IMF forms sometimes in use, the generalized Rosin–Rammmler function (Eq. *Ch* in ► *Table 4-3*, thick short-dash-dotted curve) best represents the data, apart from a deviation for $m \gtrsim 10 M_\odot$, which can be fixed by adopting a Salpeter/Massey power-law extension for

■ Table 4-2

The mass in stars, M_{ecl} , normalized to $1 M_{\odot}$ for the canonical IMF between 0.07 and $150 M_{\odot}$ in comparison to the mass for other mass ranges (m_1 – m_2) and the commonly used single power-law Salpeter IMF taken here to have $\alpha = 2.3$ and 2.35 . All IMFs are normalized to have identical $M_{0.5,150}$ (☛ 4.8) for $m > 0.5 M_{\odot}$, and no stellar remnants are included. N is the correspondingly normalised number of stars

m_1/M_{\odot}	m_2/M_{\odot}	N	M_{ecl}/M_{\odot}	IMF used
0.07	150	1.000	1.000	Canonical IMF
0.1	150	0.823	0.973	Canonical IMF
0.1	0.8	0.702	0.370	Canonical IMF
0.5	150	0.223	0.719	Canonical IMF
0.07	150	2.874	1.424	Salpeter IMF, $\alpha = 2.30$
0.1	150	1.808	1.264	Salpeter IMF
0.1	0.8	1.687	0.660	Salpeter IMF
0.5	150	0.223	0.719	Salpeter IMF
0.07	150	3.378	1.543	Salpeter IMF, $\alpha = 2.35$
0.1	150	2.087	1.348	Salpeter IMF
0.1	0.8	1.961	0.756	Salpeter IMF
0.5	150	0.238	0.719	Salpeter IMF

$m > 1 M_{\odot}$ (☛ 4.56). Interpretation of m_o in terms of a characteristic stellar mass poses difficulties. As can be seen in ☛ Fig. 4-24, the difference between a Chabrier IMF and a canonical IMF is negligible. The popular Miller–Scalo log-normal IMF (Eq. *MS* in ☛ Table 4-3) deviates strongly from the empirical data at high masses. Larson’s Eq. *Lb* in ☛ Table 4-3 fits rather well, except that it may predict too many BDs. Finally, the *effective initial mass function for galactic disks* proposed by Hollenbach et al. (2005) and Parravano et al. (2011) (Eq. *Holl*) reproduces the data in the alpha plot quite well (Fig. 1 in Hollenbach et al. 2005; Parravano et al. 2011) and is not incorporated into ☛ Fig. 4-26 here. Note, however, that a *composite IMF* (i.e., the IGIMF and the local IGIMF, LIGIMF), which would be the correct IMF for a whole disk galaxy or parts thereof, respectively, ought to be steeper (have a larger α) at high masses which is precisely what Scalo (1986) deduced for the MW disk ($\alpha_3 \approx 2.7$, (☛ 4.59), ☛ Sect. 13.1).

The closed functional IMF formulations (Eqs. *MS*, *La*, *Lb*, *Ch*, *Holl*) have the advantage that possible variations of the IMF with physical conditions can be studied more naturally than with a multi-power-law form because they typically have a characteristic mass that can be varied explicitly. However, they cannot readily capture the variation of the stellar IMF with ρ and $[\text{Fe}/\text{H}]$ discovered recently (☛ 4.65).

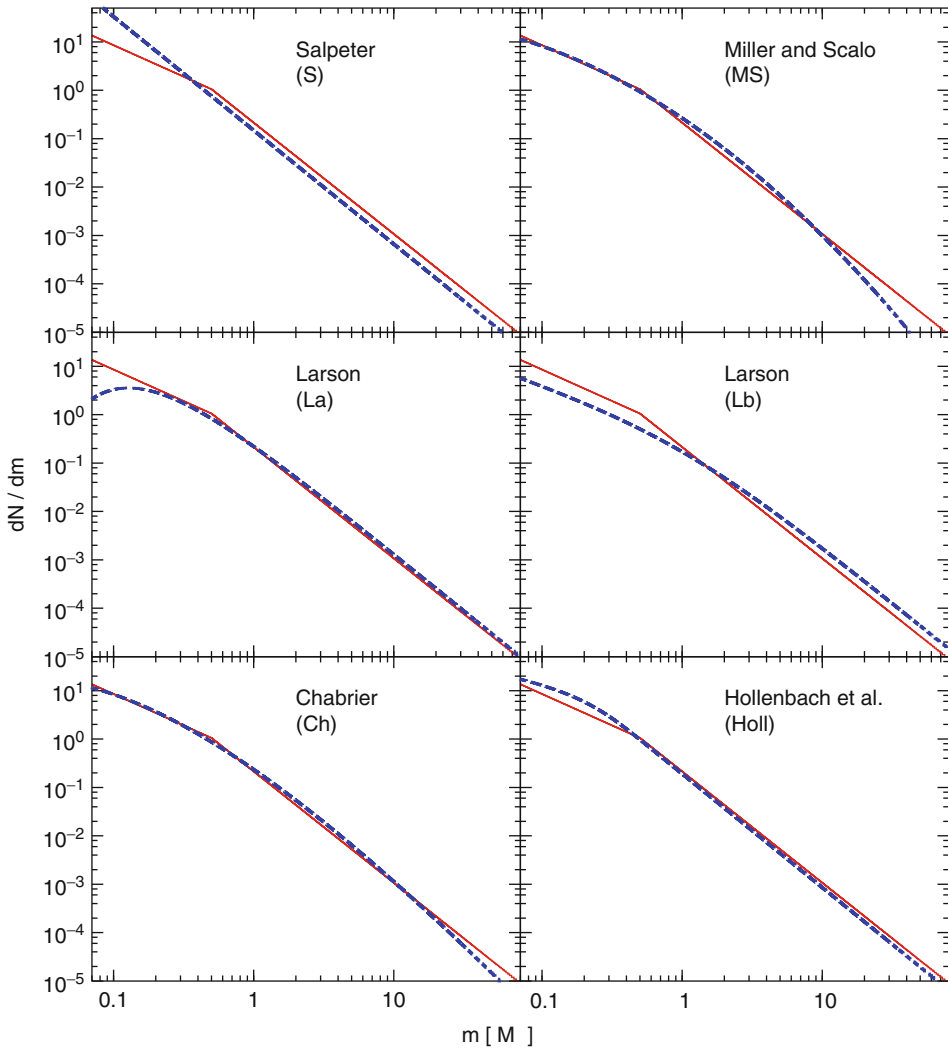
11 The Origin of the IMF

The two fundamental theoretical ansatzes for understanding the form of the IMF are discussed in ☛ Sect. 1.4, and here, a few more detailed aspects of the theoretical problem are raised. Observationally, it appears that the form of the stellar IMF is already established in the pre-stellar cloud core MF. Open questions concerning our understanding of the IMF remain.

■ Table 4-3

Summary of different proposed analytical IMF forms discussed in **Sect. 10.2** (the modern power-law form, the canonical IMF, is presented in **Eq. 4.55**), and its log-normal equivalent is given by **Eq. 4.56**). Notation: $lm \equiv \log_{10}(m/M_{\odot}) = \ln(m/M_{\odot})/\ln 10$; dN is the number of all stars in the mass interval m to $m + dm$ and in the logarithmic-mass interval lm to $lm + dlm$. The mass-dependent IMF indices, $\Gamma(m)$ (Eq. **Eq. 4.26**), are plotted in **Fig. 4-26** using the line types defined here. Equation **MS** was derived by Miller and Scalo assuming a constant star-formation rate and a Galactic disk age of 12 Gyr (the uncertainty of which is indicated in the lower panel of **Fig. 4-26**). Larson (1998) does not fit his forms (Eqs. **La** and **Lb**) to solar-neighborhood star-count data but rather uses these to discuss general aspects of likely systematic IMF evolution; the m_0 in Eqs. **La** and **Lb** given here are approximate eyeball fits to the canonical IMF

General	$dN = \xi(lm) dm = \xi_L(m) dlm$ $\xi_L(m) = (m \ln 10) \xi(m)$	gen	e.g., for power-law form: $\xi_L = A m^{\Gamma} = A m^{-\alpha}$ $\xi = A' m^{-\alpha} = A' m^{+\gamma}$ $A' = A/\ln 10$	ind
Scalo's IMF index (Scalo 1986)	$\Gamma = -\alpha = 1 + \gamma = 1 - \alpha$	Gam		
Salpeter (1955)	$\xi_L(lm) = A m^{\Gamma}$		$\Gamma = -1.35$ ($\alpha = 2.35$)	S
Miller and Scalo (1979)	$A = 0.03 \text{ pc}^{-3} \log_{10}^{-1} M_{\odot}$; $0.4 \leq m/M_{\odot} \leq 10$			
<i>thick long-dash-dotted line</i>	$\xi_L(lm) = A \exp\left[-\frac{(lm-lm_0)^2}{2\sigma_{lm}^2}\right]$		$\Gamma(lm) = -\frac{(lm-lm_0)}{\sigma_{lm}} \log_{10} e$	MS
Larson (1998)	$A = 106 \text{ pc}^{-2} \log_{10}^{-1} M_{\odot}$; $lm_0 = -1.02$; $\sigma_{lm} = 0.68$			
<i>thin short-dashed line</i>	$\xi_L(lm) = A m^{-1.35} \exp\left[\frac{-m_0}{m}\right]$ $A = -$; $m_0 = 0.3 M_{\odot}$		$\Gamma(lm) = -1.35 + \frac{m_0}{m}$	La
Larson (1998)	$\xi_L(lm) = A \left[1 + \frac{m}{m_0}\right]^{-1.35}$		$\Gamma(lm) = -1.35 \left(1 + \frac{m_0}{m}\right)^{-1}$	Lb
<i>thin long-dashed line</i>	$A = -$; $m_0 = 1 M_{\odot}$			
Chabrier (2001, 2002)	$\xi(m) = A m^{-\delta} \exp\left[-\left(\frac{m_0}{m}\right)^{\beta}\right]$		$\Gamma(lm) = 1 - \delta + \beta \left(\frac{m_0}{m}\right)^{\beta}$	Ch
<i>thick short-dash-dotted line</i>	$A = 3.0 \text{ pc}^{-3} M_{\odot}^{-1}$; $m_0 = 716.4 M_{\odot}$; $\delta = 3.3$; $\beta = 0.25$			
Hollenbach et al. (2005) and Parravano et al. (2011)	$\xi_L(m) = k m^{-\Gamma} \left[1 - \exp\left[-(m/m_{\text{ch}})^{\gamma+1}\right]\right]$			Holl
<i>not plotted in Fig. 4-26</i>	$\gamma = 0.4, \Gamma = 1.35, m_{\text{ch}} = 0.18 M_{\odot}$			

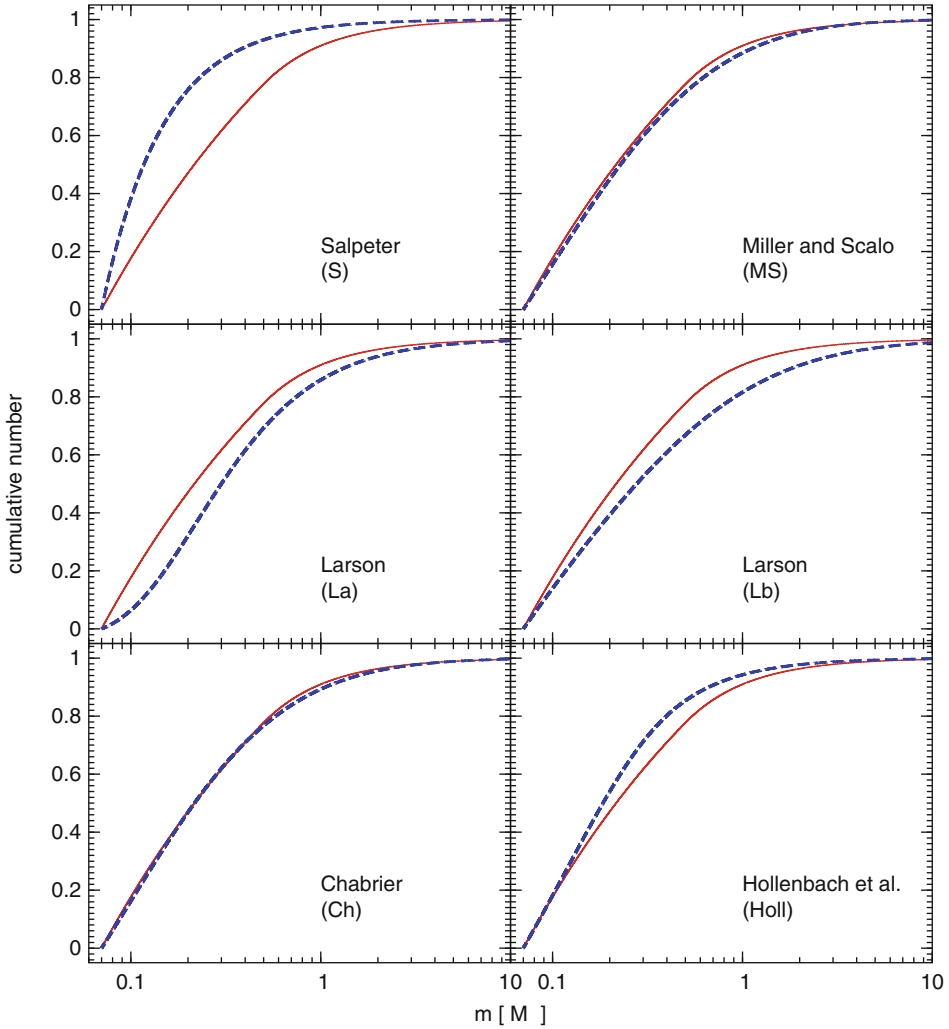


■ Fig. 4-28

A comparison between the canonical IMF (([4.55](#)), *solid red curve*) and the IMFs of [Table 4-3](#) in the interval from 0.07 to $80 M_{\odot}$. Plotted is the number of stars per mass interval versus the stellar mass, both scales being logarithmic. All IMFs are normalized such that $\int_{0.07}^{150} m \xi(m) dm = 1 M_{\odot}$, where m is the mass in solar units. Note that the “Chabrier” IMF in the *lower left panel* is the earlier Chabrier IMF listed in [Table 4-3](#) and is not the later log-normal plus power-law extension plotted in [Fig. 4-24](#). See also [Figs. 4-29](#) and [4-30](#) for a comparison of the cumulative functions

11.1 Theoretical Notions

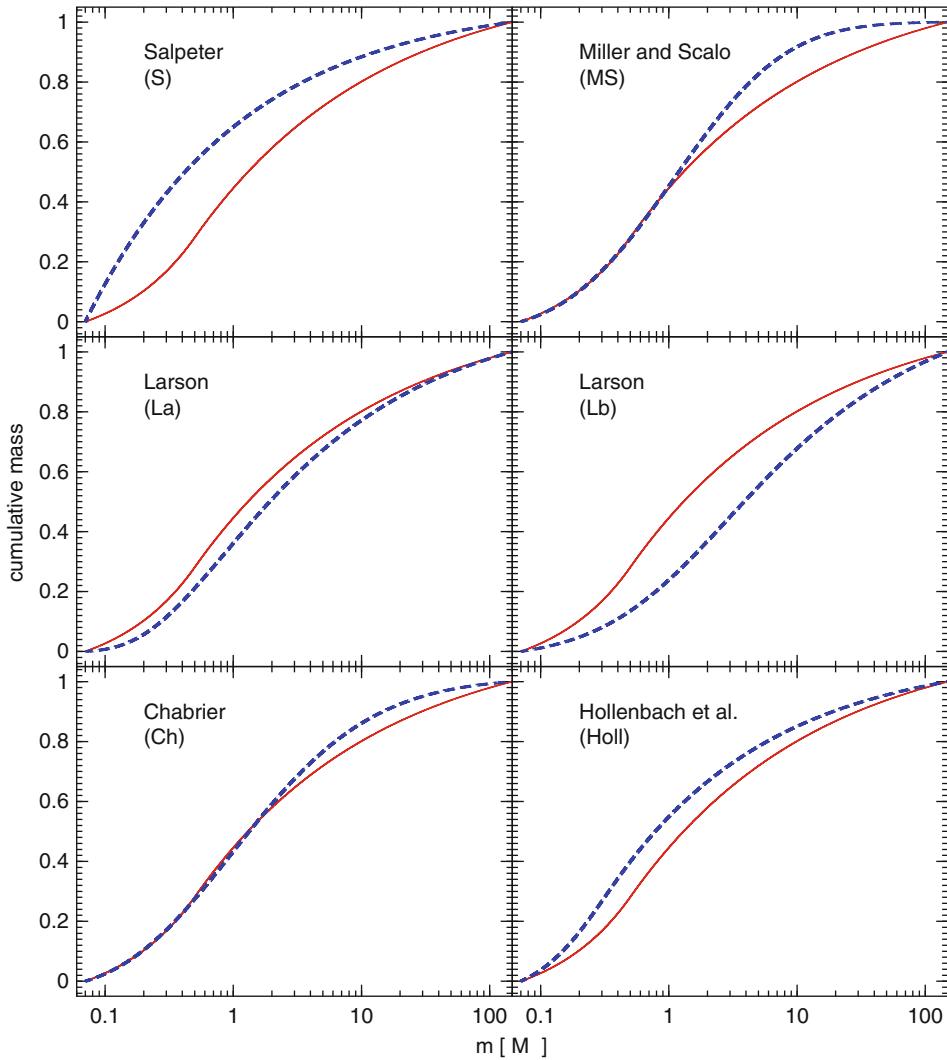
The Jeans mass scale (([4.50](#)), see also [Sect. 1.4](#)) is useful as a general indication of the rough mass scale where fragmentation of a contracting gas cloud occurs. But the concept breaks



■ Fig. 4-29

Cumulative function of the number of stars in the stellar mass range $0.07\text{--}150 M_{\odot}$, plotted here over the range $0.07\text{--}10 M_{\odot}$, for the alternative IMF forms listed in ▶ [Table 4-3](#) and plotted in ▶ [Fig. 4-28](#). In all the panels, the canonical IMF (▶ [4.55](#)) is shown as the *solid red curve*, while the alternative IMF forms are plotted with the *dashed blue curve*. Note that half of a saturated canonical stellar population has $m < 0.21 M_{\odot}$

down when considering the stellar masses that form in star clusters. The central regions of these are denser, formally leading to smaller Jeans masses which is the opposite of the observed trend, where even in very young clusters massive stars tend to be located in the inner regions. More complex physics is clearly involved (self-regulatory ansatz in ▶ [Sect. 1.4](#)). Murray and Lin (1996) develop an N -body model for the formation of stars in star clusters by considering



■ Fig. 4-30

Cumulative function of the mass in stars in the stellar mass range $0.07\text{--}150 M_{\odot}$ for the alternative IMF forms listed in [Table 4-3](#) and plotted in [Fig. 4-28](#). In all the panels, the canonical IMF ([4.55](#)) is shown as the *solid red curve*, while the alternative IMF forms are plotted with the *dashed blue curve*. Note that half the mass of a saturated canonical stellar population has $m < 1 M_{\odot}$

the formation and dynamical evolution and interactions of cloudlets which can form stars which heat their surroundings thereby quenching further star formation. They naturally arrive at mass-segregated clusters with power-law IMFs. Interestingly, Elmegreen et al. (2008) show that the Jeans mass depends only weakly on the ambient radiation field, temperature, metallicity, and density for a large range of initial conditions, therewith possibly explaining a largely invariant IMF. The insensitivity of the IMF to the initial kinetic structure of the gas is also

explicitly demonstrated by Bate (2009) in a high-resolution SPH computation of a large-scale molecular cloud. The inclusion of radiation and magnetic fields is now becoming possible and is touched upon further down in this section.

The impressive agreement between the theoretical $m_{\max} - M_{\text{ecl}}$ relation (► 4.49), which results from SPH and FLASH computations of star formation with and without feedback, and the observational data (► Fig. 4-5), together with these simulations also reproducing the general form of the IMF quite well, suggests that the essential physics of star formation has been understood. Another success of star formation computations is the excellent reproduction of the observed mass distribution of BDs and VLMSs (► Sect. 8.4).

But, as discussed in ► Sect. 2.6, this theoretical work has not yet explained the birth binary properties. Also, as stated in ► Sect. 1.4, this theoretical work predicts a variation of the IMF with the temperature and the density of the gas. The consensus in the community has been until now that such variation has not been found (more on this in ► Sect. 12.9).

Stars almost certainly regulate their own mass by feedback (winds, radiation, outflows) limiting to a certain degree the amount of mass that can be accreted (self-regulatory ansatz in ► Sect. 1.4). Indeed, this is an attractive proposition in-line with the structure in the $m_{\max} - M_{\text{ecl}}$ relation at $m_{\max} \gtrsim 10 M_{\odot}$ (► Sect. 3.3). The $m_{\max} - M_{\text{ecl}}$ relation suggests a self-regulated balance between the amount of matter that can accrete onto a star and the feedback and local fragmentation which modulates the accretion. In particular noteworthy is the result (► Sect. 3.3.3) that feedback from the central massive stars may reduce the accretion onto the surrounding less massive stars leading to an improved agreement with the observational $m_{\max} - M_{\text{ecl}}$ data. Self-regulatory behavior may be a reason why the shape of the IMF appears to be so invariant to metallicity (Krumholz et al. 2010; Myers et al. 2011).

The coagulation of protostars probably plays a significant role in the densest regions where the cloud-core collapse time, τ_{coll} , is longer than the fragment collision time scale, τ_{cr} . The collapse of a fragment to a protostar, with about 95% of its final mass, takes no longer than $\tau_{\text{coll}} \approx 0.1 \text{ Myr}$ (Wuchterl and Klessen 2001), so that the core crossing time

$$\tau_{\text{cr}}/\text{Myr} \approx 42 \left(\frac{(R/\text{pc})^3}{(M_{\text{ecl}}/M_{\odot})} \right)^{\frac{1}{2}} < 0.1 \text{ Myr}, \quad (4.60)$$

where R is the half-mass radius of a Plummer-sphere model, implies $M/R^3 > 10^5 M_{\odot}/\text{pc}^3$. Such densities correspond to star-formation rate densities

$$SFRD > 0.1 M_{\odot} \text{ year}^{-1} \text{ pc}^{-3} \quad (4.61)$$

if the embedded cluster forms over a time scale of 1 Myr. It is under such conditions that proto-stellar interactions are expected to affect the emerging stellar mass spectrum. These are found in the centers of very dense and rich embedded star clusters before they expand as a result of gas expulsion (see also Elmegreen and Shadmehri 2003 for a more detailed discussion). Thus, only for massive stars ($m \gtrsim 10 M_{\odot}$) is the form of the IMF likely affected by coagulation (Bonnell et al. 1998; Klessen 2001; Zinnecker and Yorke 2007). It is indeed remarkable that evidence for a top-heavy IMF has only emerged in star-forming systems in which $\rho_{\text{gas+stars}} > 10^5 M_{\odot} \text{ pc}^{-3}$ (► Sect. 12.8).

But the observed mass segregation in very young clusters cannot as yet be taken as conclusive evidence for primordial mass segregation that results naturally from competitive accretion/fragmentation induced starvation (p. 150) and coagulation, unless precise N -body

computations of the embedded cluster are performed for each case in question. For example, models of the ONC show that the degree of observed mass segregation can be established dynamically within about 2 Myr (see [Fig. 4-20](#)) despite the embedded and much denser configuration having no initial mass segregation. The notion behind the “no initial mass-segregation” assumption is that star clusters fragment heavily subclustered (Megeath et al. 1996; Bontemps et al. 2001; Klessen 2001; Maschberger et al. 2010; Maschberger and Clarke 2011), and each subcluster may form a few OB stars with a few hundred associated lower-mass stars ([Table 4-1](#)), so that the overall morphology may resemble a system without significant initial mass segregation (McMillan et al. 2007; Fellhauer et al. 2009) or even with “inverse mass segregation” (Parker et al. 2011).

The theoretical time scale, $t_2 - t_1$ in [Fig. 4-20](#), for mass segregation to be established even in an initially not subclustered cluster can be shortened by decreasing the relaxation time. This can be achieved by reducing the number of stars and subclusters or by increasing the pre-gas expulsion density in the model. But it may prove impossible to find agreement at the same time with the density profile and kinematics because the ONC is probably expanding rapidly now given that it is virtually void of gas (Kroupa et al. 2001). Clearly, the issue of initial mass segregation requires more study.

An interesting approach to explain why the stellar IMF has a Salpeter/Massey slope above $1 M_\odot$ has been proposed by Oey (2011) by applying the concepts of the IGIMF theory ([Sect. 13.1](#)) to the notion that a star cluster forms as a hierarchical fractal structure (Elmegreen 1997, 1999). Sally Oey is able to demonstrate that the true stellar IMF of the stars that form within each sub-clump (here referred to as the “*elementary IMF*,” EIMF) has $\alpha_{\text{scl}} \approx 2$ in agreement with the MF of embedded clusters, $\beta \approx 2$ ([4.67](#)). If the least-massive sub-clump has a mass comparable to the lowest-mass star, then the stellar IMF of the whole cluster steepens to the canonical Salpeter/Massey value, $\alpha = 2.3$ ([4.55](#)), upon addition of all the sub-clumps. Notable is that this effect had already been discovered in SPH simulations of star formation (Maschberger et al. 2010). The reason for this is that low-mass sub-clumps cannot form massive stars, which therewith become underrepresented in the whole cluster. A slightly top-heavy IMF with $\alpha = \alpha_{\text{scl}}$ may result in a star-burst cluster if all sub-clumps are more massive than the most massive star can be, since then each sub-clump can form the whole range of stellar masses, that is, if each sub-clump was saturated. This is essentially the exact same reason why the IGIMF may become top-heavy if the least massive cluster is saturated (p. 148).

The Oey model raises a number of questions: On the one hand, it remains to be seen how the other well-studied effects of coagulation, competitive accretion, and stellar feedback would affect the resultant stellar IMF, on the other hand, it leaves the question open why the pre-stellar cloud core MF already has the shape of the stellar IMF ([Sect. 11.2](#)). Last not least observations, while very difficult, would need to confirm that the stellar IMF in sub-clumps has $\alpha_{\text{sc}} \approx 2$ rather than the canonical $\alpha = 2.3$.

But what about radiation feedback and magnetic fields? The results discussed above are based nearly exclusively on pure gravo-hydrodynamical computations only, and radiation and magnetic fields are likely to significantly affect the fragmentation and accretion behavior of a gas cloud. Such calculations are highly demanding (e.g., Stamatellos et al. 2007b), and not many results allowing statistically significant statements exist yet. The fragmentation of a magnetized cluster-forming cloud clump and the properties of the emerging bound cores such as their masses, radii, mean densities, angular momenta, spins, magnetizations, and mass-to-flux ratios are documented and are found to be comparable to observational properties. Price and Bate (2009) present SPH calculations of a $50 M_\odot$ gas cloud with radiative feedback and a magnetic

field being incorporated together for the first time. They show that the magnetic field supports the gas cloud on large scales while the radiation field limits fragmentation on a small scale through heating of the gas nearby protostars. The overall result is a similar mass distribution of protostars but a significantly reduced star-formation efficiency. Li et al. (2010) find that, besides reducing the overall star formation rate, magnetic fields and outflow feedback reduce the characteristic mass of the cluster stars. The influence of a magnetic field and radiation feedback on the formation of massive stars in a rotating $1,000 M_{\odot}$ cloud is studied for the first time computationally by Peters et al. (2011a) using the adaptive-mesh FLASH code. In comparison to the otherwise equal purely gravo-hydrodynamical computations, it is found that by yielding large-scale support, the magnetic field limits secondary fragmentation and by carrying angular momentum outward, it enhances accretion onto the central massive protostar. The number of fragments is only reduced by a factor of about two though.

At extreme star-formation rate densities (SFRDs), the gas can be heated by supernova(SN)-generated cosmic rays (CRs) which propagate into the dense cloud region heating it if the SN frequency is sufficiently high to sustain a CR flux. Many SNe in confined places may also contribute to turbulent heating of the gas. Such a very excited gas phase has been revealed by high-J CO line observations of ultraluminous infrared galaxies (Papadopoulos, private communication). These physical processes raise the Jeans mass possibly leading to a top-heavy IMF (► Sect. 12.4).

11.2 The IMF from the Cloud-Core Mass Function?

The possible origin of most stellar masses is indicated by a remarkable discovery by the team around Philippe André in Saclay for the low-mass ρ Oph cluster in which star formation is ongoing. Here, the pre-stellar and protostar MF are indistinguishable, and both are startlingly similar to the canonical IMF, even in showing the same flattening of the power law near $0.5 M_{\odot}$ (Motte et al. 1998; Bontemps et al. 2001). The pre-stellar cores have sizes and densities that can be interpreted as Jeans instabilities for the conditions in the ρ Oph cloud, so that cloud fragmentation due to the collapse of density fluctuations in a dissipating turbulent interstellar medium (Nordlund and Padoan 2003; Padoan and Nordlund 2002, 2004; Mac Low and Klessen 2004; Tilley and Pudritz 2004; Hennebelle and Chabrier 2008, 2009; Elmegreen 2011) augmented with mass-dependent accretion and accretion-truncation prescription (Veltchev et al. 2011; Myers 2011, see also ► Sect. 1.4) appears to be the most-important mechanism shaping the stellar IMF for masses $0.1 \lesssim m/M_{\odot} \lesssim$ a few M_{\odot} .

If this result can be generalized, then it may indicate that the shape of the IMF describing the vast population of stars would be mostly determined by the spectrum of density fluctuations in molecular clouds. Similar results have indeed been obtained for the Serpens clouds and for the clouds in Taurus-Auriga (Testi and Sargent 1998; Onishi et al. 2002, respectively).

More recently, Alves et al. (2007) used a novel technique by mapping the extinction through the Pipe Nebula to derive the pre-stellar cloud-core MF. The result is a cloud-core MF with the same shape as the canonical IMF but shifted to larger masses by a factor of about three. And, the massive effort within the Gould Belt Survey with the Herschel telescope toward the Aquila rift and Polaris Flare regions yields virtually the same result (André et al. 2010). This is interpreted to mean that the star-formation efficiency (SFE) is about 30% independently of

the core mass. It is interesting that this SFE is similar to the efficiencies deduced for embedded clusters on a scale of about 1 pc (Lada and Lada 2003). Such a SFE also arises from magneto-hydrodynamical computations of the formation of a protostar from a cloud core: As shown by Machida and Matsumoto (2011), 20–50% of the infalling material is expelled poleward by a magnetically driven wind. Why the more recent results indicate a SFE shift in mass between the cloud-core MF and the stellar IMF which was not seen in the earlier work will need to be understood.

The majority of stellar masses making up the canonical IMF thus do not appear to suffer significant subsequent modifications such as competitive accretion, proto-stellar mergers, or even self-limitation through feedback processes, assuming there is a one-to-one map between a pre-stellar core mass and the stellar mass. A keynote result supporting the “no-interaction” conjecture is presented by André et al. (2007) using kinematical measurements in ρ Oph: The relative velocities of the pre-stellar and proto-stellar cloud cores are too small for the individual condensations to interact with each other before evolving into pre-main sequence stars. This implies the following conjecture.

The IMF Origin Conjecture

The IMF is mostly determined by cloud fragmentation at the pre-stellar stage. Competitive accretion is not the dominant mechanism at the proto-stellar stage. Competitive accretion may govern the growth of starless, self-gravitating condensations initially produced by gravo-turbulent fragmentation toward an IMF-like mass spectrum (André et al. 2007).

The work of Padoan and Nordlund (2002) with modifications by Elmegreen (2011) demonstrates that, under certain reasonable assumptions, the mass function of gravitationally unstable cloud cores deriving from the power spectrum of a supersonic turbulent medium leads to the observed canonical IMF above $1 M_{\odot}$. The flattening at lower masses is a result of a reduction of the SFE because at small masses, only the densest cores can survive sufficiently long to collapse, this being the reason why BDs do not, as a rule, form as stars do directly from self-induced gravitational collapse of a core (► Sect. 8.4). Tilley and Pudritz (2007) apply their magneto-hydrodynamical simulations to test the Padoan–Nordlund model of turbulent fragmentation finding disagreements in the position of the peak mass of the core mass spectrum.

If this holds for more massive star-forming regions is not clear. Bontemps et al. (2010) show by observation that massive young dense cores in Cygnus X tend to be more fragmented than expected from turbulence regulated monolithic collapse but that they are less fragmented than predicted by gravo-turbulent scenarios and that the fragments do not show a canonical IMF distribution. The magneto-hydrodynamical plus radiation feedback simulations discussed above would be relevant in understanding this observational result. Hennebelle and Chabrier (2008, 2009) calculate the mass spectrum of the self-gravitating cores based on an extension of the Press–Shechter statistical formalism thereby also reproducing well the observed IMF. But it depends on the turbulence spectrum and the IMF ought to therefore be variable. The generation of turbulence in the ISM through the passage of curved shock waves is investigated by Kevlahan and Pudritz (2009). They conclude that “the composite nature of the IMF—a log-normal plus power-law distribution—is shown to be a natural consequence of shock interaction and feedback from the most massive stars that form in most regions of star formation in the

galaxy.” A three-component IMF was also suggested by Elmegreen (2004) to account for the apparently seen different IMFs in different star-forming regimes.

The Similarity Statement

The intriguing observational result is that the stellar, proto-stellar, and pre-stellar clump mass spectra are similar in shape to the stellar IMF.

This is consistent with the independent finding that the properties of binary systems in the Galactic field can be understood if most stars formed in modest, ρ Oph-type clusters with primordial binary properties as observed in Taurus-Auriga (([◆ 4.47](#)), [◆ Sect. 2.6](#)) and with the independent result derived from an analysis of the distribution of local star clusters that most stars appear to stem from such modest clusters (Adams and Myers 2001). However, the canonical IMF is also similar to the IMF in the ONC ([◆ Fig. 4-19](#)) implying that fragmentation of the pre-cluster cloud proceeded similarly there.

The above similarity statement may not be entirely true because each cloud core typically forms a multiple stellar system ([◆ Sect. 2.6](#), Goodwin et al. 2008; Smith et al. 2009). In the computer simulations, the similarity of the clump mass function and the IMF does not necessarily imply a direct one-to-one mapping of clumps to stars. The temporal evolution of the clump and stellar MF in SPH models shows that stars come from a broad range of clump masses despite the similar shape of the MF.

The SPH computations by Bonnell and Bate (2002) and collaborators of the formation of dense clusters indeed not only predict the observed $m_{\max} - M_{\text{ecl}}$ relation ([◆ Fig. 4-5](#)), but they also show that a Salpeter/Massey power-law IMF is obtained as a result of competitive accretion and the merging of protostars near the cluster core driven by gas-accretion onto it, independently of metallicity as long as Z/Z_{\odot} is larger than 10^{-5} (Clark et al. 2009). The reason as to why the IMF is so invariant above a few M_{\odot} may thus be that the various physical processes all conspire to give the same overall scale-free result (see also [◆ Sect. 11.1](#)).

Open question IV emerges here: Various theories of the IMF as resulting from the pre-stellar cloud core MF account for the observed shape of the canonical IMF. This is also the case for theories based on competitive accretion and on coagulation. However, if star formation is intrinsically hierarchical through first the emergence of sub-clumps of stars that merge dynamically, then as discussed in [◆ Sect. 11.1](#), the stellar IMF in each sub-clump must be flatter with $\alpha_{\text{sc}} \approx 2$ than the Salpeter/Massey index. Why does theory not predict this flatter *elementary IMF* (EIMF)?

Open Question IV

The hierarchical model of star-cluster formation implies the EIMF to be flatter than the canonical IMF. Why has theory never predicted this EIMF?

A word of caution is advisable in view of the modeling of star formation and the resulting IMF. An excellent example of how state-of-the-art modelling may be somewhat misleading is as follows: Observations found a top-heavy PDMF in the Arches cluster with an apparent lack of

stars below $6 M_{\odot}$ (Stolte et al. 2002), and this was often interpreted as a top-heavy IMF because of the youth of the object. Klessen et al. (2007) therefore presented a state-of-the-art SPH model of star formation from warm gas in the Galactic Center which produces a top-heavy IMF with a downturn below about $6 M_{\odot}$. But Kim et al. (2006) showed that the observed Arches PDMF is in fact readily explained by strong stellar dynamical evolution due to the extreme environment with no need for a noncanonical IMF. And, stars less massive than $6 M_{\odot}$ have formed in the cluster without a sign of a deficit.

Main Result

Observations have led to the understanding that the pre-stellar cloud-core MF is very similar to the proto-stellar MF and to the stellar IMF suggesting that gravitationally driven instabilities in a turbulent medium may be the primary physical mechanism setting the shape of the IMF for stars in the mass range $0.1 \lesssim m/M_{\odot} \lesssim \text{few}$. Theoretical work has progressed significantly but remains too inconsistent with observations to allow the conclusion that a theory of the IMF exists.

12 Variation of the IMF

From the previous discussion, it has emerged that the IMF appears to be universal in resolved star-forming systems as are found largely in the vicinity of the Sun and in very nearby extragalactic systems (LMC, SMC, dSph satellites).

But the stellar IMF has been predicted theoretically to systematically vary with star-forming conditions. This has been shown with Jeans-mass arguments including SPH simulations and for self-regulated mass-growth physics (☛ Sect. 1.4). Stellar populations formed from triggered star formation in expanding shells have also been suggested to be significantly top-heavy (Dale et al. 2011).

In ☛ Sect. 1.4, the variable IMF prediction is emphasized as a robust result of IMF theory. Is there evidence supporting this prediction? Next, some recently emerging observational evidence for a dependency of the IMF on star-forming conditions is presented which may be part of the long-expected violation of the invariant IMF hypothesis (☛ Sect. 1.4).

12.1 Trivial IMF Variation Through the $m_{\max} - M_{\text{ecl}}$ Relation

The existence of the $m_{\max} - M_{\text{ecl}}$ relation (☛ Sect. 3.3) trivially implies that the stellar IMF varies with increasing stellar mass, M_{ecl} , of the population formed in the star-formation event. This is best seen by the increase of the average stellar mass with increasing M_{ecl} in contrast to what is expected if the IMF were merely a probabilistic distribution function (☛ Fig. 4-3).

Note that the relatively small scatter of the observational $m_{\max} - M_{\text{ecl}}$ data (☛ Fig. 4-5) and the sharpness of the distribution of IMF power-law indices (☛ Fig. 4-27) may be taken to imply that pure random sampling from the IMF is excluded as a viable model for stellar populations (cf. ☛ Figs. 4-1–4-3). This is further supported by the lack of evidence for massive stars forming in isolation (☛ Sect. 4) and the lack of stars more massive than a few M_{\odot} in the Taurus-Auriga and Orion Nebula star-forming regions (p. 231).

12.2 Variation with Metallicity

Differences in the metallicity, Z , of the populations sampled in the nearby Local Group do not lead to discernible variations of the IMF for massive stars as has been shown using star counts with spectroscopic classification (● Fig. 4-8). Thus, the distribution of masses for massive stars has been interpreted to not be significantly affected by the metallicity of the star-forming gas.

That low-mass stars are forming together with massive stars in the low-metallicity environment within the Magellanic Clouds with a MF similar to the canonical IMF has been demonstrated through the deep photometric surveying effort by Dimitrios Gouliermis and collaborators (e.g., Gouliermis et al. 2006; Da Rio et al. 2009). Detailed studies of star formation in the low-metallicity environment of the Small Magellanic Cloud is being pushed by this team (e.g., Gouliermis et al. 2010), but the work is very challenging due to biases through crowding, resolution, and mass estimation from photometric data. Further, Yasui et al. (2008) find there to be no measurable difference in system MFs down to $0.1 M_{\odot}$ between the extreme outer Galactic disk and the inner, more metal-rich regions. More metals lead to a dustier gas cloud and how the characteristic dust grain size may affect the final characteristic stellar mass has been studied by Casuso and Beckman (2012).

A metallicity effect for low-mass stars may, however, be uncoverable from a detailed analysis of Milky Way star clusters. Present-day star-forming clouds typically have somewhat higher metal abundances ($[\text{Fe}/\text{H}] \approx +0.2$) compared to 5 Gyr ago ($[\text{Fe}/\text{H}] \approx -0.3$) (Binney and Merrifield 1998) which is the mean age of the population defining the canonical IMF. The data in the empirical alpha plot (● Fig. 4-26) indicate that some of the younger clusters may have an individual-star MF that is somewhat steeper (larger α_1) than the canonical IMF *when unresolved binary stars are corrected for*. This may mean that clouds with a larger $[\text{Fe}/\text{H}]$ produce relatively more low-mass stars which is tentatively supported by the typically but not significantly flatter MFs in globular clusters (Piotto and Zoccali 1999) that have $[\text{Fe}/\text{H}] \approx -1.5$, and the suggestion that the old and metal-poor ($[\text{Fe}/\text{H}] = -0.6$) thick-disk population also has a flatter MF below $0.3 M_{\odot}$ with $\alpha_1 \approx 0.5$ (Reyl e and Robin 2001). If such a systematic effect is present, then for $m \lesssim 0.7 M_{\odot}$ and to first order,

$$\alpha \approx 1.3 + \Delta\alpha [\text{Fe}/\text{H}], \quad (4.62)$$

with $\Delta\alpha \approx 0.5$ (Kroupa 2001b).

Is this evidence supporting the variable IMF prediction (● Sect. 1.4)? At the present, (● 4.62) needs to be taken as suggestive rather than conclusive evidence. Measuring the stellar IMF for low-mass stars in metal-poor environments, such as in young star clusters in the Small Magellanic Cloud (Gouliermis et al. 2005, 2006; Schmalzl et al. 2008; Da Rio et al. 2009), would thus be an important goal. In ● Sect. 12.7, we will uncover somewhat more robust evidence for a variation of the stellar IMF toward top-heaviness with decreasing metallicity but coupled to increasing density, possibly yielding a for the first time uncovered variation of the overall form of the IMF with metallicity in (● 4.63) (Marks et al. 2012).

That metallicity does play a role is becoming increasingly evident in the planetary-mass regime in that the detected exo-planets appear to occur mostly around stars that are more metal-rich than the Sun (Zucker and Mazeh 2001; Vogt et al. 2002).

12.3 Cosmological Evidence for IMF Variation

Larson (1998) invoked a *bottom-light* IMF that is increasingly deficient in low-mass stars the earlier the stellar population formed, while for high-mass stars, it is equal to the canonical IMF at all times. This theoretical suggestion is motivated by the decreasing ambient temperatures due to the expansion of the Universe, implying a decrease of the Jeans mass in a star-forming gas cloud and thus a decrease of the average mass of the stars with decreasing redshift. Such an IMF could explain the relative paucity of metal-poor G-dwarfs in the solar neighborhood (the G-dwarf problem), compared to the predictions of a closed-box model for the chemical evolution of galaxies. This is because, in this model, low-mass stars form less frequently at times when the self-enrichment of galaxies has not yet reached the current level.

Empirical evidence for a bottom-light IMF variation was suggested by van Dokkum (2008) in order to explain how the integrated colors of massive elliptical (E) galaxies change with redshift. The stellar populations of E galaxies are old, so that they have evolved passively most of the time until the present day. Corrected for redshift, the stellar populations of E galaxies are therefore bluer the more distant (i.e., younger) they are. van Dokkum (2008) finds, however, that the observed reddening is less with decreasing redshift than can be accounted for by stellar evolution. This trend may be understood if the IMF were a bottom light power law with a characteristic mass $m_C \approx 2 M_\odot$ and an index near $1 M_\odot$ of $\alpha_2 \approx 1.3$ when E galaxies formed, rather than the canonical $m_C \approx 0.1 M_\odot$, $\alpha_2 = 2.3$.

But such an IMF appears to be in contradiction with the observed near-canonical PDMFs of globular clusters (GCs). Just like E galaxies, GCs have formed their stellar populations at high redshifts. Nevertheless, GCs still have a large population of stars that should not have formed in them, according to the model proposed in van Dokkum (2008). Thus, the IMF either did not have the time dependency suggested in van Dokkum (2008), or the IMF in GCs was considerably different from the one in E galaxies (see the review by Bastian et al. 2010). By studying various integrated gravity-sensitive features in luminous E galaxies, Cenarro et al. (2003) and van Dokkum and Conroy (2010) on the other hand find evidence for a very bottom-heavy IMF. Cenarro et al. (2003) propose the IMF to vary according to $\alpha = 3.41 + 2.78[\text{Fe}/\text{H}] - 3.79[\text{Fe}/\text{H}]^2$.

At present, it is unclear how these discrepant results may be brought into agreement. The detailed dynamical analysis by Deason et al. (2011) demonstrates that E galaxies are consistent with a canonical IMF rather than with a bottom-light IMF over a large range of masses. This modeling assumes dark matter to provide the mass deficit, and it is unclear whether an entirely different approach not relying on the existence of dynamically relevant dark matter would lead to a different conclusion. A contradicting result is obtained by Grillo and Gobat (2010) who show that for a sample of 13 E galaxies, a match is found between the dynamical masses and photometric (plus dark matter) masses if the IMF was a Salpeter power law, that is, if it was bottom-heavy relative to a canonical IMF (☉ Table 4-2). Chemo-evolutionary population synthesis models (Vazdekis et al. 1996, 1997) need an initially top-heavy IMF which, after a short burst of star formation, becomes bottom heavy to explain the optical and near-infrared colours and line indices of the most metal rich E galaxies. It is unclear at this stage why the dynamical and stellar population modeling of E galaxies leads to such diverging results.

Other empirical evidence for a time variability of the IMF was proposed by Baugh et al. (2005) and Nagashima et al. (2005). Baugh et al. (2005) modeled the abundances of Lyman-break galaxies and submillimeter galaxies. Both types of galaxies are considered to be distant

star-forming galaxies, but the latter are thought to be obscured by dust that transforms the ultraviolet radiation from massive stars into infrared radiation. The model from Baugh et al. (2005) is based on galaxy formation and evolution via accretion and merging according to Λ CDM cosmology. They include a detailed treatment of how the radiation from stars is converted to dust emission. Their model only returns the correct abundances of Lyman-break galaxies and submillimeter galaxies if they assume two modes of star formation. One of the two modes suggested by Baugh et al. (2005) occurs with the canonical IMF and the other one is with a *top-heavy* IMF. The mode with the top-heavy IMF is thought to be active during star bursts, which are triggered by galaxy mergers. Quiescent star formation in the time between mergers is thought to be in the mode with the canonical IMF. Nagashima et al. (2005) used the same model for galaxy evolution in order to explain the element abundances in the intra-cluster medium of galaxy clusters. They also need the two mentioned modes of star formation in order to succeed. Noteworthy is that this two-mode IMF model is qualitatively in natural agreement with the IGIMF theory (☛ Sect. 13.1). In the IGIMF theory, the IGIMF becomes top-heavy only in star bursts.

Wilkins et al. (2008) note that the stellar mass density observed in the local universe is significantly smaller than what would be expected by integrating the cosmic star-formation history within the standard (dark-matter dominated) cosmological model. They show that a single top-heavy power-law IMF with high-mass star index $\alpha = 2.15$, that is, a smaller number of long-lived low-mass stars per massive star, can reproduce the observed present-day stellar mass density.

This evidence for IMF variations stands and falls with the validity of Λ CDM cosmology. Modeling galaxy formation and evolution to be consistent with observations is in fact a major problem of Λ CDM cosmology. There are discrepancies in the Local Group (Kroupa et al. 2010; Kroupa 2012) and in the Local Volume of galaxies (Peebles and Nusser 2010) that already now shed major doubt as to the physical applicability of the concordance cosmological model to the real world. For instance, the actual galaxy population is less diverse than the one Λ CDM theory predicts (Disney et al. 2008). It thus remains to be seen whether the results by Baugh et al. (2005), Nagashima et al. (2005), and Wilkins et al. (2008) can be confirmed using a refined model for galaxy evolution or if instead a fundamentally different cosmological model, which would yield different redshift-time-distance relations, is required. It is noteworthy in this context that evidence for top-heavy IMFs in high-star formation-rate density environments has emerged, as is discussed in the following sections.

12.4 Top-Heavy IMF in Starbursting Gas

We have seen above that the IMF is largely insensitive to star-forming conditions as found in the present-day Local Group. The observed embedded clusters in these “normal” photon-dominated star-forming regions have gas plus star densities $\rho \lesssim 10^5 M_\odot \text{pc}^{-3}$ ($M_{\text{ecl}} \lesssim 10^{5.5} M_\odot$ within a half-mass radius $r_h \approx 0.5 \text{pc}$). The objects form within about 1 Myr leading to a star-formation rate density $SFRD \lesssim 0.1 M_\odot \text{pc}^{-3} \text{year}^{-1}$ for which no significant IMF variation has been found, apart from the trivial variation through the $m_{\text{max}} - M_{\text{ecl}}$ relation (☛ Sect. 12.1).

When the SFRD becomes very high, understanding the formation of individual protostars becomes a challenge, as they are likely to coalesce before they can individually collapse (☛ 4.60) such that probably a top-heavy IMF may emerge (Dabringhausen et al. 2010).

For high ambient SFRDs and a sufficiently long duration of the star burst, Papadopoulos (2010) has shown that the cosmic rays (CRs) generated by exploding supernovae of type II (SNII) heat the clouds which are too optically thick to cool therewith becoming CR dominated regions within which the conditions for star formation are significantly altered when compared to normal photon-dominated star-forming regions. This Papadopoulos-CR mechanism raises the Jeans mass and must lead to top-heavy IMFs in star-bursts (Papadopoulos et al. 2011).

This is also found by Hocuk and Spaans (2011) who compute the IMF which arises from star formation in a $800 M_{\odot}$ cloud being irradiated with X-rays and CRs as well as UV photons from a population of massive stars and an accreting supermassive black hole (SMBH) at 10 pc distance. CRs penetrate deep into the cloud therewith heating it, while the ambient X-rays lead to a thermal compression of the cloud. Their adaptive-mesh refinement computations with the FLASH code include shear which opposes gravitational contraction through self-gravity of the cloud. The turnover mass in the IMF increases by a factor of 2.3, and the high-mass index becomes $\alpha \lesssim 2$ such that the resultant IMF is top-heavy. Shear lessens the effects of the CRs and X-rays by the IMF becoming bottom heavy but only for the most massive SMBHs. Low-mass SMBHs ($< 10^6 M_{\odot}$) or star bursts without massive BHs would thus lead to top-heavy IMFs as long as the CR flux is significant.

This may well be the dominant physical mechanism why star bursts have a top-heavy IMF. Indeed, Klessen et al. (2007) and Hocuk and Spaans (2011) demonstrate, respectively, with SPH and FLASH simulations that under warmer conditions a stellar mass-spectrum forms which is dominated by massive stars. The Jeans mass can also be raised through turbulent heating through expanding supernova shells if the explosion rate of SNII is large enough, but details need to be worked out (Papadopoulos, 2011, private communication).

Observational evidence for top-heavy IMFs in regions of high SFRD has now emerged for GCs (🔗 Sect. 12.7) and UCDs (🔗 Sect. 12.8). Top-heavy IMFs are also reported in the central regions of Arp 220 and Arp 299 assuming SFRs of duration longer than at least a few 0.1 Myr (the evidence becomes less significant for a top-heavy IMF if the objects experienced a star-burst of \lesssim few 10^5 year duration). The central region of Arp 220 has a star-formation rate of $100 M_{\odot}/\text{year}$, and the large number of observed type II supernovae requires a top-heavy IMF (Lonsdale et al. 2006; Parra et al. 2007, see also 🔗 Sect. 12.8). Noteworthy is that the Arp 220 central region consists of what may be UCD-sized substructures (🔗 Sect. 12.8). The high frequency and spatially confined occurrence of supernovae in Arp 299 indicates that the star formation occurs in highly sub-clustered regions with dimensions less than 0.4 pc and within larger structures of 30 pc scale (Ulvestad 2009). This is reminiscent of the cluster-complex model for the origin of some UCDs (Fellhauer and Kroupa 2002; Brüns et al. 2011). Likewise, Pérez-Torres et al. (2009) and Anderson et al. (2011) deduce a top-heavy IMF in Arp 299 from the relative frequency of different supernova types.

12.5 Top-Heavy IMF in the Galactic Center

Observations of the Galactic Center revealed one or two disks of about 6-Myr-old stars orbiting the central supermassive black hole (SMBH) at a distance between 0.04 and 0.4 pc. Nayakshin and Sunyaev (2005) argued that the X-ray luminosity of the Sgr A* field is too low to account for the number of young $< 3 M_{\odot}$ stars expected from a canonical IMF, considering the large number of O stars observed in the disks. The low X-ray luminosity may be explained by a low-mass cutoff near $1 M_{\odot}$. A more recent systematic search of OB stars in the central parsec revealed

a significant deficit of B-type stars in the regime of the young disks, suggesting a strongly top-heavy IMF for these disks with a best-fit power law of $\alpha_3 = 0.45 \pm 0.3$. (Bartko et al. 2010). This appears to be the very best evidence for a truly top-heavy IMF.

Using SPH simulations of star formation in fragmenting gas accretion disks, Bonnell and Rice (2008) find that the IMF of the disk stars can be bimodal (and thus top-heavy) if the infalling gas cloud is massive enough ($\approx 10^5 M_\odot$) and the impact parameter of the encounter with the SMBH is as small as ≈ 0.1 pc. Thus, only under quite extreme conditions will a top-heavy IMF emerge, whereby the strong tidal forces and the rotational shear seem to be the dominant physical mechanisms shaping the IMF since only cloud cores massive enough can collapse to a star. This is explicitly calculated by Hocuk and Spaans (2011) for a range of SMBH masses (► Sect. 12.4).

Observations of the central parsec of the Milky Way show that this region is dominated by a dense population of old stars with a total mass of $\approx 1.5 \times 10^6 M_\odot$. This stellar cluster around the SMBH has also been probed for evidence for a non canonical IMF. By means of stellar evolution models using different codes, Löckmann et al. (2010) show that the observed luminosity in the central parsec is too high to be explained by a long-standing top-heavy IMF as suggested by other authors, considering the limited amount of mass inferred from stellar kinematics in this region. In contrast, continuous star formation over the Galaxy's lifetime following a canonical IMF results in a mass-to-light ratio and a total mass of stellar black holes (SBHs) consistent with the observations. Furthermore, these SBHs migrate toward the center due to dynamical friction, turning the cusp of visible stars into a core as observed in the Galactic center. It is thus possible to simultaneously explain the luminosity and dynamical mass of the central cluster and both the presence and extent of the observed core since the number of SBHs expected from a canonical IMF is just enough to make up for the missing luminous mass.

In conclusion, observations of the Galactic center are well consistent with continuous star formation following the canonical IMF. Only the centermost young stellar disks between about 0.04 and 0.4 pc from the SMBH show a highly top-heavy IMF, but the circumstances that led to their formation must be very rare since these have not affected most of the central cluster.

12.6 Top-Heavy IMF in Some Star-Burst Clusters

There are indications of top-heavy IMFs such as in some massive star-burst clusters in the M82 galaxy. Using spectroscopy of the unresolved M82-F cluster, Smith and Gallagher (2001) derive, via the inferred velocity dispersion, a mass and from the luminosity a mass-to-light ratio that is significantly smaller than the ratio expected from the canonical IMF for a 60 Myr population. The implication is that the M82-F population may be significantly depleted in low-mass stars, or equivalently it may have a top-heavy IMF, provided the velocity dispersion is representative of the entire cluster. A possibility that will have to be addressed using stellar-dynamical modeling of forming star clusters is that M82-F may have lost low-mass stars due to tidal shocking (Smith and Gallagher 2001). Highly pronounced mass segregation which leads to a dynamically decoupled central core of OB stars is an important mechanism for reducing the measured mass-to-light ratio (Boily et al. 2005), while rapid expulsion of residual gas from forming clusters enhances the measured mass-to-light ratios (Goodwin and Bastian 2006). But also a younger age would reduce the inferred depletion in low-mass stars, and some hints exist that M82-F might be as young as 15 Myr (McCraday et al. 2005).

In an extensive literature study of Galactic and extragalactic observations, Elmegreen (2005) concluded that dense star-forming regions like star bursts might have a slightly shallower IMF, a view shared by Eisenhauer (2001).

12.7 Top-Heavy IMF in Some Globular Clusters (GCs)

Observations of 17 globular clusters (GCs) for which PDMFs were measurable over the mass range $0.5\text{--}0.8 M_{\odot}$, showed the higher-metallicity GCs to have flatter PDMFs (Djorgovski et al. 1993). This lacked an explanation until now. In particular, this trend is difficult to reconcile with standard dynamical evolution scenarios as it is unclear how dynamics could possibly know about the metal content of a cluster.

De Marchi et al. (2007) performed a deep homogeneous star-count survey of 20 Milky Way GCs using the HST and VLT and measured the global PDMFs in the mass range $0.3\text{--}0.8 M_{\odot}$. They discovered the least concentrated GCs to have a bottom-light PDMF, while the other GCs show a canonical MF. N -body calculations predict that two-body-encounter-driven dynamical evolution preferentially removes low-mass stars from a star cluster (Vesperini and Heggie 1997; Baumgardt and Makino 2003). This occurs because two-body relaxation drives the cluster into core collapse and energy conservation leads to the expansion and thus evaporation of low-mass stars. Thus, the most concentrated star clusters are expected to show the strongest depletion of low-mass stars, in disagreement with the observations.

A possibility would be that the low-concentration GCs were formed with a bottom-light IMF. However, there is no known theory of star formation which could account for this: The low-concentration clusters typically have a higher metallicity than the high-concentration clusters (see below in this section), and so the data would imply that the IMF ought to have been bottom-light in the higher-metallicity GCs. This, however, contradicts basic star-formation theory (► Sect. 1.4).

This apparent disagreement between theory and observation can be resolved if GCs formed mass-segregated and with the canonical IMF (► 4.55) for stars less massive than about $1 M_{\odot}$. Baumgardt et al. (2008) performed N -body models using the Aarseth code and demonstrate that the range of PDMFs observed by De Marchi et al. (2007) can be arrived at if GCs were born mass segregated and filling their tidal radii such that they do not need to first evolve into core collapse and so they immediately begin losing low-mass stars. However, the stellar dynamically induced trend of PDMF with concentration is not able to account for the observed metallicity dependence. Furthermore, it is not clear why GCs ought to be formed mass-segregated but filling their tidal radii since all known young clusters are well within their tidal radii.

An alternative scenario by Marks et al. (2008) also assumes the young compact GCs to be formed mass segregated with a canonical IMF below $1 M_{\odot}$. But after formation, the expulsion of residual gas unbinds the low-mass stars that typically reside near the outer region of the clusters, leading to flattening of the MF. This ansatz allows for a metal dependency, since the process of residual-gas expulsion is expected to be enhanced for metal-richer gas which has a stronger coupling to radiation than metal-poor gas, similar to the metallicity-dependent stellar winds. For each of the 20 GCs in the De Marchi et al. (2007) sample, the best-fitting tidal-field strength, radius, star-formation efficiency, and gas-expulsion time scale are obtained. This uncovers remarkable correlations between the gas-expulsion quantities, the tidal field strength, and the metallicities, allowing a very detailed reconstruction of the first collapse phase of the Milky Way about 12 Gyr ago (Marks and Kroupa 2010). The correlations, for example, confirm

the expectation that gas expulsion is more efficient and thus dynamically more damaging in metal-richer gas and also that metal-poorer GCs were denser than their metal-richer slightly younger counterparts which were subject to stronger tidal fields.

This in-turn suggests that in order to provide enough feedback energy to blow out the residual gas, the IMF had to be top-heavy in dependence of the initial density of the GC (Marks et al. 2012). Assuming the gas leaves a cluster with the velocity of the sound speed of about 10 km/s, the gas-expulsion time scales for clusters with radii between 0.5 and 1 pc would lie between $\tau_{\text{gas}} = 0.05$ and 0.1 Myr. The resultant high-mass IMF slopes derived for the GCs from their individual PDMFs assuming $0.05 \lesssim \tau_{\text{gas}}/\text{Myr} \lesssim 0.15$ cover a wide range, $0.9 \lesssim \alpha_3 \lesssim 2.3$, where α_3 is the slope for $m \gtrsim 1 M_{\odot}$, for an IMF that is canonical otherwise (► Fig. 4-31).

The calculated IMF slopes also correlate with the metallicity of the GCs, such that the PDMFs show the observed correlation after the metallicity-dependent gas-expulsion process ends and the remaining GC revirializes. This correlation with metallicity, quantified in (► 4.63) (► Fig. 4-32), is an important clue, as it implies that GCs formed from metal-poorer gas were more compact and had a more top-heavy IMF, just as is indeed expected from star-formation theory (► Sects. 1.4 and ► 11.1).

The Top-Heavy Stellar IMF/Metallicity Dependence

The suggested dependence of α_3 on globular cluster-forming cloud metallicity can be parametrized as

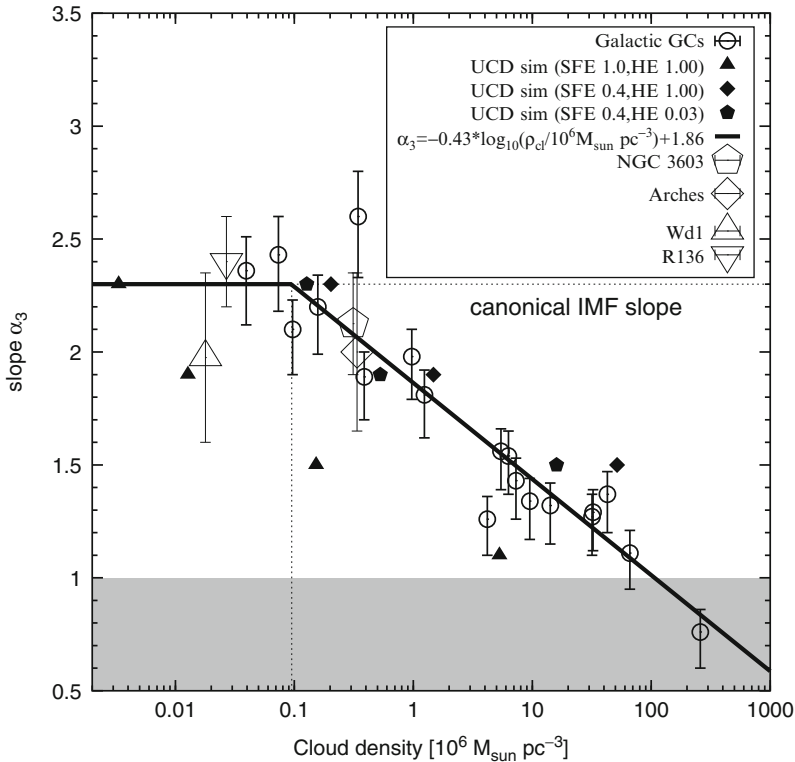
$$\begin{aligned} \alpha_3 &= \alpha_2, & m > 1 M_{\odot} \wedge [\text{Fe}/\text{H}] \geq -0.5, \\ \alpha_3 &= 0.66 [\text{Fe}/\text{H}] + 2.63, & m > 1 M_{\odot} \wedge [\text{Fe}/\text{H}] < -0.5. \end{aligned} \quad (4.63)$$

Strader et al. (2011) observed high-resolution spectra of 200 GCs of the Andromeda galaxy and discovered that the near-infrared M/L ratios decrease significantly with increasing metallicity of the GCs. This cannot be explained by secular dynamical evolution but follows from a PDMF which is systematically more bottom-light for more metal-rich GCs. This is thus the same finding as discussed above for the 20 GCs of the MW and a possible explanation put forward by Strader et al. (2011) is metallicity-dependent gas expulsion.

Marks et al. (2012) find that α_3 decreases with increasing pre-globular cluster cloud-core density, ρ ((► 4.64), ► Fig. 4-31). Such a trend is to be expected theoretically if the massive and initially dense GCs, each of which was a star burst (star-formation rate density $\text{SFRD} \gtrsim 0.1 M_{\odot}/(\text{year pc}^3)$ for an initial half-mass radius of about 0.5 pc and formation time scale of about 1 Myr), involves the merging of proto-stellar cores since the collision probability is higher in denser systems (► 4.60).

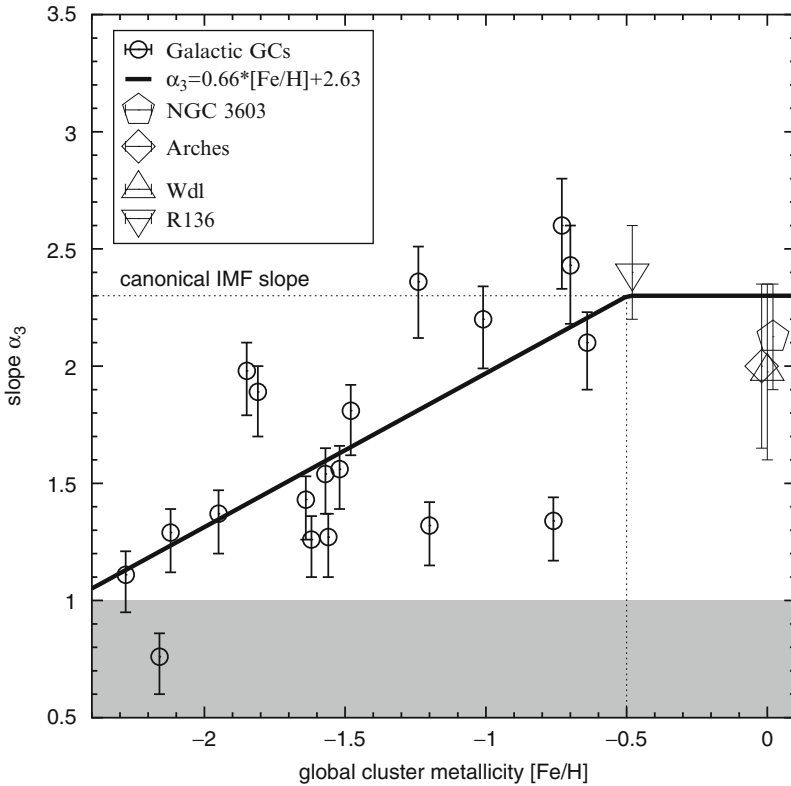
The high-mass IMF index α_3 thus depends on ρ and on $[\text{Fe}/\text{H}]$. The increasing top-heaviness, deduced with a principle-component-type analysis by Marks et al. (2012), with decreasing $[\text{Fe}/\text{H}]$ and increasing ρ is quantified here for the first time in (► 4.65) and ► Fig. 4-34. It is remarkable that the theoretically expected trend of the IMF with ρ and $[\text{Fe}/\text{H}]$ has now emerged from elaborate stellar-dynamical analysis of deep observations of GCs.

The extremely top-heavy IMFs for some of the GCs raise the question whether they could survive the strong mass loss these IMFs imply due to stellar evolution. This is especially an issue if clusters start mass-segregated (Vesperini et al. 2009).



■ Fig. 4-31

The high-mass IMF slope, α_3 , deduced for Galactic GCs (open circles) as a function of the pre-GC cloud-core density within their initial half-mass radii (● Sect. 12.7). Upper limits are for a gas-expulsion time scale of $\tau_{\text{gas}} = 0.15$ Myr, open circles are for $\tau_{\text{gas}} = 0.1$ Myr, and the lower limits are for $\tau_{\text{gas}} = 0.05$ Myr. Also plotted are MF slope values found in the literature for the massive, young clusters NGC 3603, Arches, Wd1 and R136. Their corresponding cloud-core densities were calculated using their PD half-masses within their PD half-mass radii (0.2pc for NGC3603, 0.24pc for Arches, 1pc for Wd1, 1.1pc for R136) assuming the clusters formed with a SFE of 1/3, that is, that their gaseous progenitors were three times more massive and that their sizes did not change. The filled symbols correspond to simulations devoted to finding the most probable IMF slopes in systems of different density that lead, after residual gas expulsion and supernova driven evolution, to objects that resemble the properties of UCDs today (● Sect. 12.8). Different symbols correspond to different input parameters (star-formation efficiency SFE and heating efficiency HE; see ● Sect. 12.8). The overall trend for GCs and UCDs is consistent in the sense that denser systems form flatter IMFs. The solid line is a fit to the GC data (● 4.64). Below $1 M_{\odot}$, the assumed IMFs are equal to the canonical IMF (● 4.5) and (● 4.55). The gray-shaded region at $\alpha_3 < 1$ are IMFs which contain more than 99% mass in stars with $m > 1 M_{\odot}$, making cluster survival after supernova explosions unlikely. Note that if the GCs and UCDs form on a time scale of 1 Myr, then their star-formation rate densities would be $0.1\text{--}100 M_{\odot}/(\text{year pc}^3)$. From Marks et al. (2012)



■ Fig. 4-32

The high-mass IMF slope, α_3 , deduced for Galactic GCs (open circles) as a function of the GC metallicity (▶ Sect. 12.7). Otherwise as ▶ Fig. 4-31

12.8 Top-Heavy IMF in UCDs

Further independently obtained evidence for top-heavy IMFs at high SFRDs comes from ultra compact dwarf galaxies (UCDs) which have been observed in nearby galaxy clusters. They typically have effective radii of a dozen pc and are understood to have formed as a star burst (SFRD $\approx 1 - 100 M_{\odot}/(\text{year pc}^3)$) for an initial radius of about 1 pc and formation time scale of 1 Myr, Dabringhausen et al. 2009). The Papadopoulos-CR-heating process (▶ Sect. 12.4) may be a factor during the formation of UCDs. They are, like galaxies, collisionless stellar-dynamical systems such that two-body relaxation-driven evaporation of low-mass stars is insignificant, contrary to the case for GCs (Anders et al. 2009; Miskeld and Hilker 2011; Forbes and Kroupa 2011).

For a significant sample of UCDs, high-resolution spectra are available allowing estimates of their stellar velocity dispersions. These velocity dispersions and the effective radii imply dynamical masses between 10^6 and $10^8 M_{\odot}$. The dynamical masses of UCDs are thus similar to the values of the much more extended dwarf spheroidal (dSph) galaxies. Combining the dynamical masses of UCDs with their luminosities leads to estimates for their dynamical mass-to-light (M/L) ratios. While the uncertainties of these estimates are large, the most likely values for these

M/L ratios are systematically higher than compared to the expectation for a stellar population that formed with the canonical IMF.

Dark matter is not a viable explanation for the enhanced M/L ratios of UCDs, given the current understanding of how UCDs are formed. One such idea is that UCDs evolve from star cluster complexes as are observed in interacting systems like the Antennae (NGC 4038 and NGC 4039, Brüns et al. 2011). The 300–500 Myr old ultra-massive “star cluster” W3 in the merger remnant galaxy NGC 7252 is indeed an object that supports such a model (Fellhauer and Kroupa 2005) since it is too young for it to be a stripped nucleus of a nucleated dwarf galaxy. Another idea is that UCDs are extremely massive GCs (e.g., Mieske et al. 2002). This notion is motivated by the similarities that UCDs share with GCs, for example, their seemingly continuous mass-radius relation. Thus, there is supportive evidence for both concepts, but they also imply that UCDs are essentially free of dark matter. Also note that dSph galaxies are usually thought to populate the least massive and therefore the most dense dark-matter halos. However, the dark-matter densities inferred from the dynamics of dSph galaxies are still about two orders of magnitude too low to influence the dynamics of UCDs to the required extent.¹⁷ This leaves a variation of the IMF as the most natural explanation for the high dynamical M/L ratios of UCDs (Dabringhausen et al. 2009).

In an old stellar system like a UCD, both a *bottom-heavy* IMF and a *top-heavy* IMF lead to a M/L ratio that exceeds the expectation for the canonical IMF. In the case of a top-heavy IMF, the M/L ratio of the stellar population is high only if it is old. This is because in an old population, the massive (and therefore bright) stars have turned into essentially nonluminous remnants. In the case of a bottom-heavy IMF, the M/L ratio of the stellar population is enhanced by the high M/L ratios of low-mass stars. The two cases are not easy to distinguish by observations simply because in either case, a population that is characterized by its low luminosity would have to be detected.

If the high M/L ratios of UCDs are caused by a bottom-heavy IMF, then this should be traceable for the highest M/L ratios by a characteristic absorption feature in the spectra of low-mass stars (Mieske and Kroupa 2008).

If, in contrast, the high M/L ratios of UCDs are the consequence of a top-heavy IMF, then this may be noticeable by the number of UCDs with bright X-ray sources. A system can form low-mass X-ray binaries (LMXBs) by stellar-dynamical encounters between black holes and neutron stars on the one hand and main-sequence stars on the other. When the main-sequence star evolves, its companion remnant may accrete its envelope therewith becoming visible as an LMXB. Assuming a canonical IMF, it has been shown that the incidence of LMXBs in GCs generally follows the expected correlation with GC mass. UCDs, however, turn out to be far overabundant as X-ray sources. This overabundance of UCDs as X-ray sources can be accounted for with the same top-heavy IMF dependence on UCD mass as is obtained independently from matching the M/L ratios. This constitutes a strong indication that UCDs formed with top-heavy IMFs (Dabringhausen et al. 2012).

The results reported in [Sect. 12.7](#) on GCs and here on UCDs are compared in the α_3 vs. birth-density diagram ([Fig. 4-31](#), [Sect. 4.64](#)). A remarkable agreement of how α_3 varies with birth density for GCs and UCDs emerges.

Would UCDs with the deduced top-heavy IMF survive mass loss through stellar evolution? Dabringhausen et al. (2010) calculate a set of numerical models for the early dynamical evolution of UCDs with the canonical IMF and top-heavy IMFs, using the particle-mesh code

¹⁷Adiabatic contraction (Blumenthal et al. 1986) in UCDs may, however, alleviate this problem but unlikely sufficiently so (Murray 2009).

SUPERBOX. They assume that UCDs are hyper-massive star clusters, that is, that their stellar population formed in a starburst that took place in a dense molecular cloud. A short formation time scale of the UCDs is indeed suggested by the enhanced α -element abundances that Evstigneeva et al. (2007) report for them. This implies that all massive stars of a UCD evolve over a time span of approximately 40 Myr, that is, the lifetime of the least massive stars that become a type II supernova.

The main driver for the dynamical evolution of a UCD during this epoch is the mass-loss through gas expulsion and supernova explosions. The rate of this mass loss depends on a number of parameters: the rate at which energy is deposited into the interstellar gas of the UCD, the rate at which the interstellar gas is replenished by the ejecta from type II supernovae, the star formation efficiency (SFE, which sets how much interstellar gas can be expelled aside from supernova ejecta), and the heating efficiency (HE, which is the fraction of the energy inserted into the interstellar medium that is not radiated away but used up by removing gas from the UCD). This mass loss and the consequences are quantified in Dabringhausen et al. (2010) by comparing the energy input by type II supernova explosions and the radiation of stars over a small time interval with the binding energy of the gas that is bound to the UCD at that time, leading to individual mass-loss histories for each of the considered UCD models. These mass-loss histories are implemented into the code that is used to calculate their dynamical evolution.

The mass loss is more pronounced the more top-heavy the IMF is because more matter is set free by stars that evolve fast and the energy deposition rates are high. As a result, UCDs with a very top-heavy IMF dissolve because of heavy mass loss, except for very high SFEs. Such high-effective SFEs may be realistic because the matter lost from stars will accumulate within the UCD and may form new stars (Pflamm-Altenburg and Kroupa 2009a; Wünsch et al. 2011). In that case, the UCD models with a high-mass IMF slope close to $\alpha_3 = 1$ evolve into objects that resemble an observed UCD at the end of the calculation. For moderate SFEs, the models in Dabringhausen et al. (2010) evolve into UCD-like objects if they have either the canonical IMF or a moderately top-heavy IMF with a high-mass slope in between 1.5 and 2. However, as shown above, the case of the canonical IMF can be excluded due to the high M/L ratios of and high LMXB occurrences in UCDs. The UCD models that are consistent with the constraints set by observed UCDs also have IMFs and initial densities that are remarkably close to the high-mass IMF slope versus initial cloud densities derived for Galactic GCs (Sect. 12.7) based on their PDMFs, as can be seen in Fig. 4-31.

The variation of the IMF with cluster-forming cloud density can be summarized as follows:

The Top-Heavy Stellar IMF/Density Dependence

Resolved stellar populations show an invariant IMF (4.55), but for $\text{SFRD} \gtrsim 0.1 M_{\odot}/(\text{year pc}^3)$, the IMF becomes top-heavy, as inferred from UCDs and some GCs. The dependence of α_3 on cluster-forming cloud density ρ (stars plus gas) can be parametrized for $m > 1 M_{\odot}$ as (Marks et al. 2012)

$$\begin{aligned} \alpha_3 &= \alpha_2, & \rho < 9.5 \times 10^4 M_{\odot}/\text{pc}^3, \\ \alpha_3 &= 1.86 - 0.43 \log_{10}(\rho_{\text{cl}}/(10^6 M_{\odot} \text{pc}^{-3})), & \rho \geq 9.5 \times 10^4 M_{\odot}/\text{pc}^3. \end{aligned} \quad (4.64)$$

This IMF is in good agreement with the supernova rate observed in Arp 220 and Arp 299 (☛ Sect. 12.4). Note that for $m \leq 1 M_{\odot}$, the IMF is canonical (☛ 4.55). Note also that the top-heavy IMF is a fit to the GC constraints and that this provides a good description of the independent constraints arrived at by the UCDS.

12.9 The Current State of Affairs Concerning IMF Variation with Density and Metallicity and Concerning Theory

The results achieved over the past decade in self-consistent gravo-hydrodynamical modeling of star formation in turbulent clouds with the SPH and FLASH methods have been very successful (☛ Sect. 3.3). This is evident in the overall reproduction of the stellar IMF as well as of the $m_{\max} - M_{\text{ecl}}$ relation allowing detailed insights into the physics driving the growth of an ensemble of stars forming together in one CSFE (see definitions on p. 126) with and without feedback. Also, computations with these same techniques of the fragmentation of circum-*proto-stellar* disks lead to an excellent agreement with the BD IMF. These simulations have not yet been able to reproduce the birth binary-star properties (☛ Sect. 2.6) which may be due to as yet necessarily inadequate inclusion of the various feedback processes but also because the smoothing length and sink-particle radius required in every SPH simulation limits the binary-orbital resolution to at least a few 100 AU. Computational star formation has also ventured into the difficult terrain of including magnetic fields and radiative feedback finding a certain degree of compensating effects in terms of the emerging stellar masses with a significantly reduced star-formation efficiency (☛ Sect. 11.1).

Following on from above, the current situation of our understanding of IMF variations may be described as follows (☛ Sects. 1.4 and ☛ 11.1): Theory has, over decades, robustly predicted the IMF to vary with star-forming conditions such that metal-poor environments and/or warmer gas ought to lead to top-heavy IMFs. Observations and their interpretation including corrections for biases have, on the other hand, been indicating the IMF to be invariant, this being the consensus reached by the community as mitigated in most reviews. The suggestions for top-heavy IMFs in star bursts (e.g., Elmegreen 2005; Eisenhauer 2001) were typically taken to be very uncertain due to the evidence stemming from distant unresolved and hard-to-observe star-forming systems.

Now theoretical work has set itself the task of explaining the invariance, and various suggestions have been made: Bate (2005), Bonnell et al. (2006), Elmegreen et al. (2008), Bate (2009), Krumholz et al. (2010), and Myers et al. (2011). But evidence in favor of IMF variations has proceeded to come forth.

Perhaps the first tentative indication from resolved stellar populations for a possible change of the IMF toward a bottom-heavy form with increasing metallicity of the star-forming gas has emerged through the analysis of recent star-forming events (☛ 4.62). This evidence is still suggestive rather than conclusive. And even if true, it is very difficult to extract any IMF variation because of combinations of the following issues that mask IMF variations as they typically act randomizingly.

Masking IMF variations

- Major uncertainties in pre-main sequence stellar evolution tracks (Footnote 3 on p. 121).
- The loss of low-mass stars from young, intermediate, and old open clusters through residual gas expulsion and secular evolution.
- Different evolutionary tracks of initially similar star clusters subject to different tidal fields.
- By observing the outcome of current or recent star formation, we are restricted to it occurring under very similar physical conditions.
- By using the canonical IMF as a benchmark, IMF variations become more difficult to unearth: the variation about the mean is smaller than the difference between the extrema.
- The fossils of star-forming events that were very different to our currently observationally accessible ones are given by Galactic GCs and dSph satellite galaxies. But given their typical distances, the PDMFs were not reliably measurable below about $0.5 M_{\odot}$. The evidence for or against variations of the MF has thus been limited to the mass range $0.5\text{--}0.8 M_{\odot}$.

The landmark paper by De Marchi et al. (2007) (↻ Sect. 12.7) for the first time provided unambiguous evidence for a systematically changing global PDMF in Galactic GCs. This breakthrough became possible because the PDMF could be measured down to about $0.3 M_{\odot}$ with the HST and VLT for a homogeneous sample of 20 GCs giving us a greater leverage on the PDMF and the dynamical history of the GCs. It emerges that the only model able to account for the observed variation and its correlation with metallicity is one in which the IMF becomes increasingly top-heavy with increasing density (↻ 4.64) and decreasing metallicity (↻ 4.63) in a gas-expulsion scenario. A gas expulsion origin is independently also suggested as an explanation for the metallicity– M/L anticorrelation which follows from a high-resolution analysis of 200 GCs of the Andromeda galaxy (Strader et al. 2011).

Furthermore, the modern observations of UCDs have led to the discovery that they typically have somewhat elevated M/L ratios which can be explained by a top-heavy IMF systematically changing with UCD mass. The enhanced frequency of UCDs with LMXBs can also be explained with a top-heavy IMF systematically changing with UCD mass. And, both independent results on how the IMF varies with UCD mass agree (↻ Sect. 12.8). Last not least, the deduced variation of the IMF in UCDs is in good agreement with the variation of the IMF as deduced from the de Marchi, Paresce, and Pulone GCs, when expressed in terms of the density of the star-forming cloud. Notwithstanding this agreement, the arrived at variation of the IMF is also consistent with the top-heavy IMF suggested in star-bursting systems such as Arp 220 and Arp 299 (↻ Sect. 12.4, ↻ 12.8). The evidence for top-heavy IMFs thus comes from GCs and UCDs that had SFRDs higher than any other known stellar-dynamical system including elliptical galaxies which had global $SFRD \lesssim 4 \times 10^{-9} M_{\odot} \text{pc}^{-3} \text{year}^{-1}$ for radii of about 5 kpc and formation time scales of about 0.5 Gyr.

The thus inferred systematic variation of the stellar IMF with metallicity is documented in ↻ Fig. 4-33. Note that there is an implicit dependence of the IMF on the density of the star-forming cloud (↻ 4.64). The metallicity and density dependencies are correlated because metal-poorer gas clouds may collapse to larger pre-cluster densities than metal-rich clouds which fragment earlier and into smaller masses.

This correlation is evident in the dependency of α_3 on ρ and $[\text{Fe}/\text{H}]$ calculated by Marks et al. (2012) using a principle component-type analysis of the GCs discussed in ↻ Sect. 12.7. It is formulated in (↻ 4.65) and plotted in ↻ Fig. 4-34.

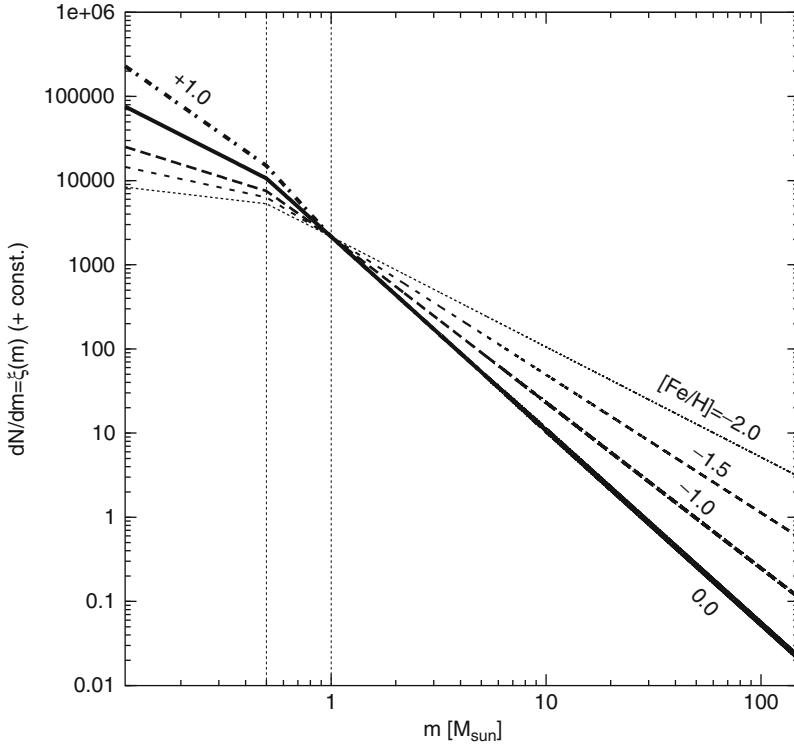


Fig. 4-33

The variation of the stellar IMF with metallicity (4.62) and (4.63). The IMFs are normalized to agree at $1 M_{\odot}$. Note that the figure is based for $m < 1 M_{\odot}$ on an extrapolation of (4.62) below about $[Fe/H] \approx -0.5$. Is this the long sought after systematic variation of the stellar IMF with metallicity? (From Marks et al. (2012))

The Stellar IMF Dependence on Density and Metallicity

Resolved stellar populations show an invariant IMF (4.55), but for $SFRD \gtrsim 0.1 M_{\odot}/(\text{year pc}^3)$, the IMF becomes top-heavy, as inferred from deep observations of GCs. The dependence of α_3 on cluster-forming cloud density, ρ , (stars plus gas) and metallicity, $[Fe/H]$, can be parametrised as

$$\begin{aligned}
 \alpha_3 &= \alpha_2, & m > 1 M_{\odot} \wedge x < -0.89, \\
 \alpha_3 &= -0.41 \times x + 1.94, & m > 1 M_{\odot} \wedge x \geq -0.89, & (4.65) \\
 x &= -0.14 [Fe/H] + 0.99 \log_{10} (\rho / (10^6 M_{\odot} \text{pc}^{-3})).
 \end{aligned}$$

How does this suggested dependence of the IMF on metallicity compare with the observational exclusion of any metallicity dependence in the Local Group (4.8)? The SMC has $[Fe/H] \approx -0.6$ such that on average $\alpha_3 \approx 2.1$ (using (4.63)) which is consistent with the SMC datum in 4.8. For the LMC $[Fe/H] \approx -0.4$ such that on average $\alpha_3 \approx 2.3$ which is consistent with the LMC data in 4.8. Thus, the IMF variation deduced from GCs and UCDs are easily accommodated by the Local Group data.

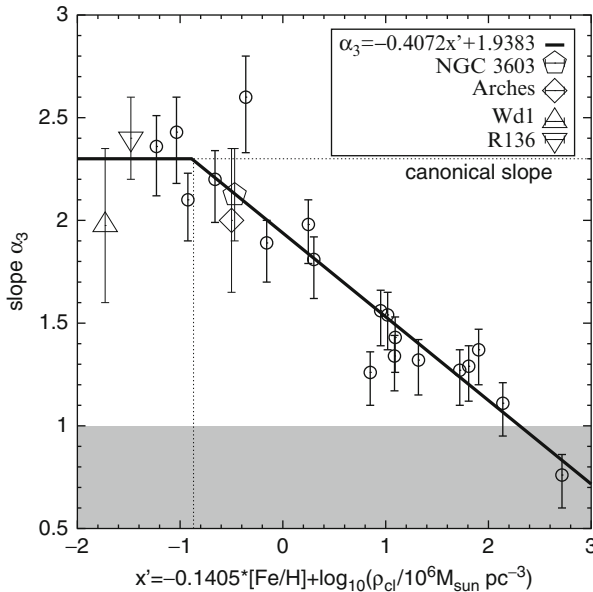


Fig. 4-34

The variation of the stellar IMF with metallicity and cloud density (4.65) as deduced from deep observations of MW GCs using a principal-component-type analysis (From Marks et al. (2012)). Otherwise as Fig. 4-31

On galaxy scales, tentative evidence for IMF variations has begun to emerge in 2003 with the prediction that the IMF in a whole galaxy ought to be steeper (top-light) if the stellar IMF were invariant (Sect. 13.1). This (unwillingly) resolved the long-standing but mostly ignored discrepancy between the canonical IMF index $\alpha = 2.3$ and the Scalo IMF index for the MW field ($\alpha \approx 2.7$), and generalization to galaxies of different type lead to an immediate understanding of the metallicity–galaxy-mass relation and other systematic effects.

The first observational evidence for a systematic change of the galaxy-wide IMF indeed appeared in 2008 and 2009 (Sect. 13.2 below).

In Conclusion of this Sect. 12

Recent research has begun to uncover increasing evidence for a possible systematic variation of the IMF in agreement with the theoretical expectations, but this evidence requires verification by further observational work. At present, the here quantified systematic variation of the IMF is more of a suggestion than established fact, but this suggestion may give a framework and a target for future work.

13 Composite Stellar Populations: The IGIMF

The IMF in individual pc(cluster)-scale star-forming events ranging in mass from a few M_{\odot} up to $10^8 M_{\odot}$ is reasonably well constrained, as the previous sections have shown. Integrated galaxy-wide properties, on the other hand, depend on the galaxy-wide content of all newly

formed stars, that is, on the composition of all collective star-formation events (CSFEs, i.e., embedded star clusters, [▶ Sect. 1](#)) in a galaxy.

Following philosophical approach A ([▶ Sect. 1.5](#)), it has usually been assumed that the galaxy-wide IMF is identical to the canonical stellar IMF which is established on the pc-scale events. This is based on the simplest-of-all assumptions that the stellar distribution is sampled purely randomly from the invariant IMF, that is, that the IMF is a probabilistic density distribution function (e.g., Elmegreen 2004). Thus, for example, 10^5 clusters, each of $10 M_\odot$, would have the same composite (i.e., combined) IMF as one cluster with mass $10^6 M_\odot$. This assumption, while being simple, has important implications for the astrophysics of galaxies. For example, luminosities such as the $H\alpha$ flux would scale linearly with the SFR leading to the much discussed Kennicutt–Schmidt star-formation law, $\Sigma_{\text{SFR}} \propto \Sigma_{\text{gas}}^N$ with $N \approx 1.5$, where Σ_{SFR} and Σ_{gas} are the SFR surface density and gas-mass surface density, respectively (Kennicutt et al. 1994; Kennicutt 2008).

The existence of the physical $m_{\text{max}} - M_{\text{ecl}}$ relation ([▶ Sect. 3.3](#)) has, on the other hand, profound consequences for *composite populations*. It immediately implies, for example, that 10^5 clusters, each weighing $10 M_\odot$, *cannot* have the same composite (i.e., combined) IMF as one cluster with $10^6 M_\odot$ because such small clusters can never make stars more massive than about $2.5 M_\odot$ ([▶ Fig. 4-5](#)). And since low-mass clusters are far more numerous than massive clusters, galaxies would have steeper composite, or integrated galactic IMFs (IGIMFs), than the stellar IMF in each individual cluster (Kroupa and Weidner 2003, also hinted at independently by Vanbeveren 1982). Furthermore, massive-star-sensitive galaxy luminosities would not scale linearly with the SFR leading to a significant revision of the $H\alpha$ –SFR relation with corresponding major implication for the galaxy-wide star-formation law ($N = 1$ instead of 1.5, [▶ 4.73](#)) below.

This is indeed supported to be the case by the theory of star formation (see the BVB conjecture on p. 148) which implies optimal sampling to possibly be closer to reality than the purely probabilistic IMF approach.

13.1 IGIMF Basics

The galaxy-wide IMF, the *integrated galactic IMF*, is the sum of all the stellar IMFs in all CSFEs formed over a time span δt . While this is a next to trivial concept, it turns out to be extremely powerful in particular when its foundation is sought in approach B ([▶ Sect. 1.5](#)), that is, in *optimal sampling* ([▶ Sect. 2.2](#)). The IGIMF is therefore the integral over the embedded cluster MF (ECMF, ξ_{ecl}):

Definition

The IGIMF is an integral over all star-formation events in a given star-formation “epoch” $t, t + \delta t$,

$$\xi_{\text{IGIMF}}(m; t) = \int_{M_{\text{ecl}, \text{min}}}^{M_{\text{ecl}, \text{max}}(\text{SFR}(t))} \xi(m \leq m_{\text{max}}(M_{\text{ecl}})) \xi_{\text{ecl}}(M_{\text{ecl}}) dM_{\text{ecl}}, \quad (4.66)$$

with the normalization conditions equations ([▶ Eqs. 4.69](#) and [▶ 4.70](#))

$$M_{\text{ecl}} - m_{\text{max}}(M_{\text{ecl}}) = \int_{0.07 M_\odot}^{m_{\text{max}}(M_{\text{ecl}})} m' \xi(m') dm',$$

$$1 = \int_{m_{\text{max}}(M_{\text{ecl}})}^{m_{\text{max}^*}} \xi(m') dm',$$

which together yield the $m_{\text{max}} - M_{\text{ecl}}$ relation ([▶ 4.11](#)).

Here, $\xi(m \leq m_{\max}) \xi_{\text{ecl}}(M_{\text{ecl}}) dM_{\text{ecl}}$ is the stellar IMF contributed by $\xi_{\text{ecl}} dM_{\text{ecl}}$ CSFEs with stellar mass in the interval $M_{\text{ecl}}, M_{\text{ecl}} + dM_{\text{ecl}}$. The ECMF is often taken to be a power law,

$$\xi_{\text{ecl}}(M_{\text{ecl}}) \propto M_{\text{ecl}}^{-\beta}, \quad (4.67)$$

with $\beta \approx 2$ (Lada and Lada 2003), whereby an “embedded cluster” is taken here to be a CSFE and not a gravitationally bound star cluster (see definitions on p. 126). $M_{\text{ecl,max}}$ follows from the maximum star-cluster mass versus global star-formation rate of the galaxy,


$$M_{\text{ecl,max}} = 8.5 \times 10^4 \left(\frac{\text{SFR}}{M_{\odot}/\text{year}} \right)^{0.75}, \quad (4.68)$$

(equation 1 in Weidner and Kroupa 2005, as derived by Weidner et al. 2004 using observed maximum star cluster masses). A relation between $M_{\text{ecl,max}}$ and SFR , which is a good description of the empirical data, can also be arrived at by resorting to optimal sampling. It follows by stating that when a galaxy has, at a time t , a $\text{SFR}(t)$ over a time span δt over which an optimally sampled embedded star cluster distribution builds up with total mass $M_{\text{tot}}(t)$, then there is one most massive CSFE,

$$1 = \int_{M_{\text{ecl,max}}(t)}^{M_{\text{U}}} \xi_{\text{ecl}}(M'_{\text{ecl}}) dM'_{\text{ecl}}, \quad (4.69)$$

with M_{U} being the physical maximum star cluster that can form (for practical purposes, $M_{\text{U}} > 10^8 M_{\odot}$), and


$$\text{SFR}(t) = \frac{M_{\text{tot}}(t)}{\delta t} = \frac{1}{\delta t} \int_{M_{\text{ecl,min}}}^{M_{\text{ecl,max}}(t)} M'_{\text{ecl}} \xi_{\text{ecl}}(M'_{\text{ecl}}) dM'_{\text{ecl}}. \quad (4.70)$$

$M_{\text{ecl,min}} = 5 M_{\odot}$ is adopted in the standard modeling and corresponds to the smallest “star-cluster” units observed (the low-mass sub-clusters in Taurus-Auriga in  Fig. 4-5).

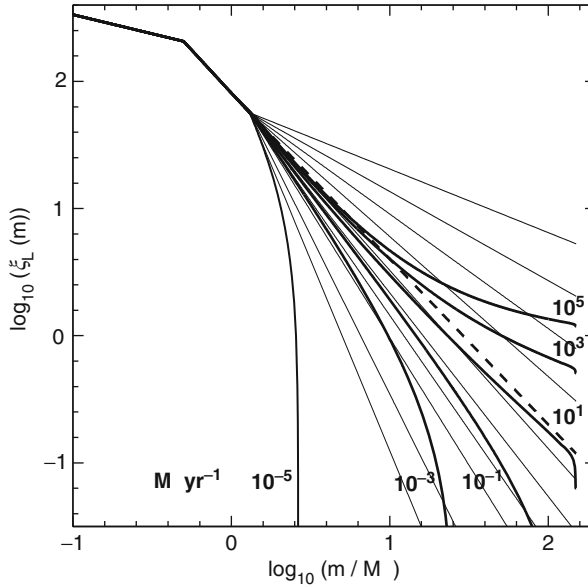
Weidner et al. (2004) define δt to be a “star-formation epoch,” within which the ECMF is sampled optimally, given a SFR. This formulation leads naturally to the observed $M_{\text{ecl,max}}(\text{SFR})$ correlation if the ECMF is invariant, $\beta \approx 2.35$ and if the “epoch” lasts about $\delta t = 10$ Myr. Thus, the embedded cluster mass function is optimally sampled in about 10 Myr intervals, independently of the SFR. This time scale is nicely consistent with the star-formation time scale in normal galactic disks measured by Egusa et al. (2004) using an entirely independent method, namely, from the offset of HII regions from the molecular clouds in spiral-wave patterns. In this view, the ISM takes about 10 Myr to transform via molecular cloud formation to a gas-free population of dispersing young simple stellar populations.

The time-integrated IGIMF then follows from

$$\xi_{\text{IGIMF}}(m) = \int_0^{\tau_{\text{G}}} \frac{\xi_{\text{IGIMF}}(m; t)}{\delta t} dt, \quad (4.71)$$

where τ_{G} is the age of the galaxy under scrutiny. The time-integrated IGIMF, $\xi_{\text{IGIMF}}(m)$, is the stellar IMF of all stars ever to have formed in a galaxy and can be used to estimate the total number of supernovae ever to have occurred, for example. $\xi_{\text{IGIMF}}(m; t)$, on the other hand, includes the time dependence through a dependency on $\text{SFR}(t)$ of a galaxy and allows one to compute the time-dependent evolution of a stellar population over the lifetime of a galaxy, for example, its instantaneous population of massive stars ( Fig. 4-35). Note that

$$\xi_{\text{IGIMF}}(m) = \left(\frac{\tau_{\text{G}}}{\delta t} \right) \xi_{\text{IGIMF}}(m; t) \quad \text{if } \text{SFR}(t) = \text{const.} \quad (4.72)$$



■ Fig. 4-35

The dependence of the logarithmic IGIMF (☛ 4.66) on the SFR of a galaxy. The IGIMF is normalized by the total number of stars such that it does not change visibly at low stellar masses in this plot. This IGIMF has been computed by adopting the canonical IMF which becomes top-heavy at embedded star-cluster densities $\rho > 10^5 M_{\odot}/(\text{year pc}^3)$ (☛ 4.64), an ECMF with $\beta = 2$, $M_{\text{ecl},\text{min}} = 5 M_{\odot}$, and the semi-analytical $m_{\text{max}} - M_{\text{ecl}}$ relation (☛ 4.12). For a given M_{ecl} , $\rho = (\frac{1}{2} M_{\text{ecl}}/SFE) 3/(4\pi r_{0.5}^3)$ is the cloud (stellar plus gas) density, whereby a star-formation efficiency of $SFE = 1/3$ and initial half-mass radius $r_{0.5} = 0.5 \text{ pc}$ are assumed (Marks and Kroupa 2010; Dabringhausen et al. 2010; Marks and Kroupa 2012). The *thin lines* are IMFs with different power-law indices, α' , for $m > 1.3 M_{\odot}$ (the IGIMF is identical to the canonical IMF (☛ 4.55) below this mass) $\alpha' = 1.5, 1.7, 1.9, 2.1, 2.3, 2.4, 2.6, 2.8, 3.0, 3.5, 4.0$ (top to bottom), whereby the canonical value $\alpha' = 2.3 = \alpha_3$ is shown as the *thick dashed line*. Thus, for example, the IGIMF has $1.9 < \alpha' < 2.1$ when $SFR = 10 M_{\odot}/\text{year}$

13.2 IGIMF Applications, Predictions and Observational Verification

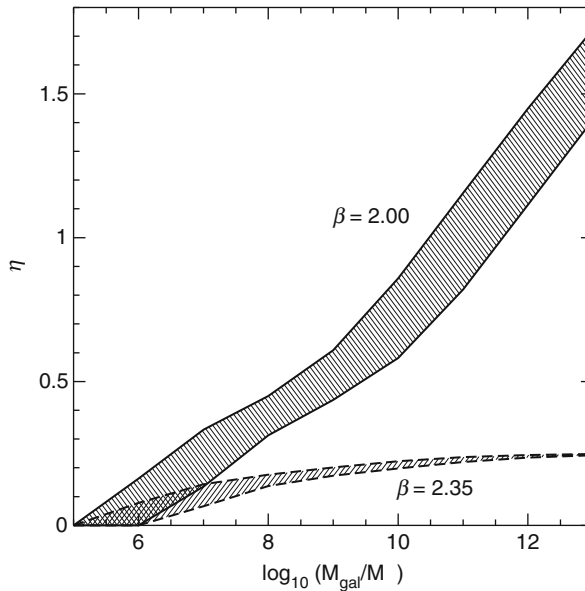
Since stellar clusters with larger masses are observed to form at higher SFRs (☛ 4.68), the ECMF is sampled to larger masses in galaxies that are experiencing high SFRs, leading to IGIMFs that are flatter than for low-mass galaxies that have only a low-level of star-formation activity. Weidner and Kroupa (2005) show that the sensitivity of the IGIMF power-law index for $m \gtrsim 1 M_{\odot}$ toward SFR variations increases with decreasing SFR. In star bursts extending to $SFR \approx 10^4 M_{\odot}/\text{year}$ the IGIMF can become top-heavy (Weidner et al. 2011) because the ultra-massive star clusters that enter the IGIMF integral have top-heavy IMFs (☛ 4.64). The dependence of the IGIMF on the SFR of a galaxy is shown in ☛ Fig. 4-35.

Thus, galaxies with a small mass in stars can either form with a very low continuous SFR (appearing today as low-surface-brightness but gas-rich galaxies) or with a brief initial SF burst (dE or dSph galaxies). It is also possible for a dwarf galaxy to evolve through multiple bursts such that its IGIMF varies. Thus, the IGIMF ought to vary significantly among dwarf galaxies (☛ Fig. 4-35). In all cases, however, the IGIMFs are invariant for $m \lesssim 1.3 M_{\odot}$ which is the maximal stellar mass in $5 M_{\odot}$ “clusters” (☛ Fig. 4-5), assuming $M_{\text{ecl,min}} = 5 M_{\odot}$ to be the invariant lower-mass limit of CSFEs.

An interesting application of the IGIMF theory to a particular system, our MW Bulge, is as follows: By studying the metallicity distribution of Bulge stars, (Ballero et al. 2007a, b among others) deduce the Bulge of the MW to have formed rapidly on a time scale of $\lesssim 1$ Gyr with a top-heavy IMF with $\alpha_3 \lesssim 2.1$. Given that the mass of the MW Bulge amounts to about $10^{10} M_{\odot}$, it follows that the Bulge would have formed with $SFR > 10 M_{\odot}/\text{year}$. From ☛ Fig. 4-35, it can be seen that the resulting IGIMF has an equivalent power-law index $1.9 < \alpha' < 2.1$ for $m > 1.3 M_{\odot}$, in excellent agreement with the IMF constraints based on the metallicity distribution. The IGIMF theory therefore naturally accounts for the Bulge IMF; that is, no parameters have to be adjusted apart from specifying the SFR.

Because the IGIMF steepens above about $1.3 M_{\odot}$ with decreasing SFR, this being the *IGIMF effect*, all galaxy-wide applications based on a constant IMF require a critical consideration and possibly a complete revision. A few studies on the outcome when a galaxy-wide constant IMF is replaced by a SFR-dependent IGIMF do already exist: Low-surface-brightness galaxies would appear chemically young, while the dispersion in chemical properties ought to be larger for dwarf galaxies than for more massive galaxies (Goodwin and Pagel 2005; Weidner and Kroupa 2005). The observed mass-metallicity relation of galaxies can be naturally explained quantitatively in the IGIMF context (Köppen et al. 2007). The $[\alpha/\text{Fe}]$ element abundance ratios of early-type galaxies, which decreases with decreasing stellar velocity dispersion, can be understood as an IGIMF effect (Recchi et al. 2009) with the associated reduction of the need for downsizing. And indeed, the chemical evolution modeling of the Fornax dwarf-spheroidal satellite galaxy demonstrates that this system must have produced stars up to at most about $25 M_{\odot}$ in agreement with the prediction of the IGIMF theory given the low $SFR \approx 3 \times 10^{-3} M_{\odot}/\text{year}$ deduced for this system when it was forming stars in the past (Tsuji-moto 2011). Another interesting implication is that the number of supernovae per star would be significantly smaller over cosmological times than predicted by an invariant Salpeter IMF (Goodwin and Pagel 2005, ☛ Fig. 4-36), except in phases when the average cosmological SFR is higher than about $10 M_{\odot}/\text{year}$.

The relation between the produced total $\text{H}\alpha$ luminosity and the underlying SFR is linear in the classical Kennicutt picture, that is, in the context of a constant galaxy-wide IMF. It turns out that this relation becomes strongly nonlinear at SFRs comparable to the SMC and smaller (Pflamm-Altenburg et al. 2007). The implication of the revised $L_{\text{H}\alpha}$ -SFR relation is fundamental: In the classical picture, the calculated gas depletion times (τ , the ratio of available neutral gas mass and current SFR) of dwarf irregular galaxies are much larger than those of large disk galaxies. This has been taken to mean that dwarf galaxies have lower “star-formation efficiencies,” $1/\tau$, than massive galaxies. But the IGIMF-revised $L_{\text{H}\alpha}$ -SFR relation reveals a fundamental constant gas depletion time scale of about $\tau = 3$ Gyr over almost five orders in magnitude in total galaxy neutral gas mass (Pflamm-Altenburg and Kroupa 2009b). Dwarf galaxies thus have “star-formation efficiencies” comparable to those of massive galaxies.



■ Fig. 4-36

The number of supernovae of type II (SNII) per star in the IGIMF divided by the number of SNII per star in the canonical IMF, η , as a function of the stellar galaxy mass, M_{gal} . The *upper shaded area* is for an ECMF with $\beta = 2$, while the *lower shaded area* assumes $\beta = 2.35$, both with $M_{\text{ecl,min}} = 5 M_{\odot}$. The *upper bound* for each *shaded region* is for an initial SF burst model of 1 Gyr duration ($SFR = M_{\text{gal}}/1 \text{ Gyr}$), while the *lower bounds* are for a constant SFR over a Hubble time ($SFR = M_{\text{gal}}/13.7 \text{ Gyr}$). For details, see Weidner and Kroupa (2005), but the here plotted η is computed from the IGIMFs shown in Fig. 4-35. For most galaxies, the SNII rate is expected to be smaller than the expected rate for an invariant IMF (Goodwin and Pagel 2005). But it can be seen that the number of SNII per star becomes larger than for a canonical IMF when $\beta = 2.0$ and $SFR > 10 M_{\odot}/\text{year}$ because the IGIMF becomes top-heavy when CSFEs with cloud density $\rho > 10^5 M_{\odot}/\text{pc}^3$ are included. Star-bursting galaxies therewith come with an overabundance of type II SN if the ECMF has $\beta = 2$ and clusters with mass down to $5 M_{\odot}$ form in star bursts. Similar results are obtained for $\beta = 2.35$ and an ECMF without low-mass clusters (not plotted here). Such and similar calculations allow observational testing of the IGIMF theory, for example, through constraining the mass of the extreme SN progenitors in dependence of the SFR of the host galaxy (Neill et al. 2011)

Furthermore, it is possible to formulate the IGIMF on local scales and not only on global (galaxy-wide) ones. This can be achieved straightforwardly by replacing all galaxy-wide quantities in the IGIMF theory (4.66) by their corresponding surface densities. This local IGIMF (LIGIMF) theory (Pflamm-Altenburg and Kroupa 2008) readily explains the observed radial $\text{H}\alpha$ cutoff in disk galaxies as well as the different radial profiles in $\text{H}\alpha$ and FUV observed by Boissier et al. (2007).

Summarizing the two star-formation laws of galaxies which emerge from the IGIMF theory (Pflamm-Altenburg and Kroupa 2007, 2008, 2009b),

$$\frac{SFR}{M_{\odot} \text{ year}^{-1}} = \frac{1}{2.8 \text{ Gyr}} \frac{M_{\text{gas}}}{M_{\odot}},$$

$$\frac{\Sigma_{\text{SFR}}}{M_{\odot} \text{ pc}^{-2} \text{ year}^{-1}} = \frac{1}{2.8 \text{ Gyr}} \frac{\Sigma_{\text{gas}}}{M_{\odot} \text{ pc}^{-2}}, \quad (4.73)$$

where SFR is the global star-formation rate of the galaxy with mass in neutral gas mass of M_{gas} and Σ_{SFR} and Σ_{gas} are the surface star-formation rate and surface gas densities, respectively.

A compilation of a number of issues relating to the astrophysics of galaxies that are naturally resolved within the IGIMF theory are listed in the box IGIMF successes (p. 230).

The IGIMF concept has allowed a number of predictions: Based on the IGIMF theory, a decreasing galaxy-wide $H\alpha$ /FUV-flux ratio with decreasing total SFR has been predicted (Pflamm-Altenburg et al. 2009). This prediction has been confirmed qualitatively (Meurer et al. 2009) and quantitatively (Lee et al. 2009). Additionally, Hoversten and Glazebrook (2008) found, in the integrated properties of over 50,000 SDSS galaxies, that galaxies of lower mass seem to have steeper IMFs than more massive ones, as would be expected from the IGIMF. A direct confirmation of the IGIMF would be to measure the IGIMF effect. A few predictions and tests are compiled in the box IGIMF predictions/tests (p. 231).

IGIMF Successes

The mass–metallicity relation of galaxies emerges naturally (Köppen et al. 2007);

The $[\alpha/\text{Fe}]$ element abundance ratios of early-type galaxies emerge naturally (Recchi et al. 2009).

The observed radial $H\alpha$ cutoff in disk galaxies as well as the different radial profiles in $H\alpha$ and FUV emerge naturally (Pflamm-Altenburg and Kroupa 2008).

The SFR of a galaxy is proportional to its mass in neutral gas (4.73).

The gas depletion time scales of dwarf irregular and large disk galaxies are about 2.8 Gyr, implying that dwarf galaxies do not have lower star-formation efficiencies than large disk galaxies (Pflamm-Altenburg and Kroupa 2009b).

The stellar-mass buildup times of dwarf and large galaxies are only in agreement with downsizing in the IGIMF context but contradict downsizing within the traditional framework that assumes a constant galaxy-wide IMF. The stellar-mass buildup times in dwarf galaxies become shorter than a Hubble time and therewith naturally solve the hitherto unsolved problem that the times are significantly longer than a Hubble time if an invariant IMF is assumed (Pflamm-Altenburg and Kroupa 2009b).

The IGIMF solution for the IMF of the Galactic Bulge is in excellent agreement with the top-heavy IMF derived from chemical-evolution studies of the Bulge.

For a $SFR = 119 M_{\odot}/\text{year}$, the IGIMF has $\alpha \approx 2$ (Fig. 4-35). This is in good agreement with the constraint $\alpha = 1.9 \pm 0.15$ observed for the $z \approx 2.5$ lensed galaxy SMM J163554.2+661225 with Herschel by Finkelstein et al. (2011), who adopted a maximum stellar mass of $100 M_{\odot}$ whereas the IGIMF theory adopts $m_{\text{max}^*} = 150 M_{\odot}$ therewith biasing their IMF solution to slightly steeper indices.

IGIMF Predictions/Tests

If the IMF were a stochastic or probabilistic distribution function then a population of a e.g., 1500 very young stars should contain 9 stars more massive than $8 M_{\odot}$ (☛ [Table 4-1](#)). The finding by Hsu et al. (2012) that the L1641 cloud is deficient in O and early B stars to a 3-4 sigma significance level constitutes a direct observational verification of the IGIMF effect (many low-density star-forming clumps).

In the young star-forming region Taurus-Auriga, random sampling from the IMF predicts 9 stars more massive than $3.25 M_{\odot}$ while none are observed. However, such a low number of stars above $3.25 M_{\odot}$ is to be expected if one assumes a local IGIMF effect from the sub-clusters in that region. The closest star-forming region is thus fully consistent with the IGIMF theory while being in conflict with the hypothesis that star formation is equivalent to randomly sampling stars from the IMF.

The fraction among all stars of massive stars in a galaxy with a low SFR is smaller than in a galaxy with a larger SFR. For example, $SFR = 10^{-3} M_{\odot}/\text{year}$, no star more massive than $18 M_{\odot}$ ought to be seen in the galaxy while 30 are expected for an invariant canonical stellar IMF (Weidner et al, in preparation).

The number of type II supernovae is smaller in all dwarf and normal galaxies than hitherto thought assuming an invariant stellar IMF (☛ [Fig. 4-36](#)).

These new insights should lead to a revision of theoretical work on galaxy formation that typically until now relied on an invariant IMF. Empirical evidence in favor of or against the notion of a galaxy-variable IGIMF is being studied (e.g., Corbelli et al. 2009; Calzetti et al. 2010; Fumagalli et al. 2011; Neill et al. 2011; Roychowdhury et al. 2011; Weisz et al. 2012) and will ultimately lead to a refinement of the ideas. Important for workers to realize here is that stellar-dynamical processes are a central physics ingredient when analyzing populations of unresolved star clusters and the spatial distribution of massive stars. Also, care needs to be exercised in testing hypotheses self-consistently (☛ [Sect. 1.6](#)). At a fundamental level, the IGIMF theory is correct since a galaxy is trivially the sum of all star-formation events. The astrophysical constraints over many orders of magnitude of galaxy mass allow one to constrain the fundamental parameters that define the particularly valid IGIMF. These parameters are the ECMF (e.g., do dwarf and massive galaxies form only massive clusters when they experience a star burst?), the exact form of the $m_{\text{max}} - M_{\text{ecf}}$ relation, and the variation of the stellar IMF with star-formation rate density.

Main Results

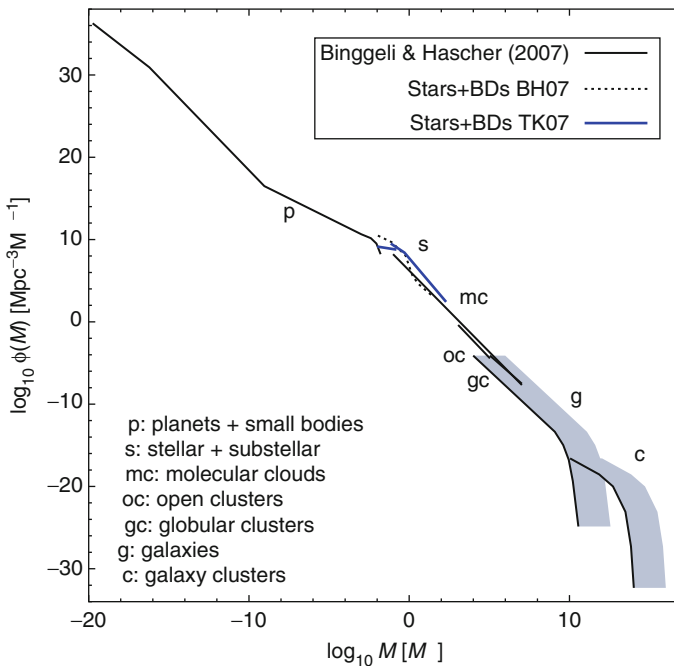
Due to the clustered nature of star formation, the composite or integrated IMF of galaxies or of parts thereof is steeper than the canonical IMF for low to modest SFRs ($SFR \lesssim 1 M_{\odot}/\text{year}$) or SFR surface densities, respectively, and it can be top-heavy for larger SFRs. The behavior of the IGIMF depends on the variation of the ECMF with the SFR. The IGIMF theory for the first time links in a computationally accessible way empirically calibrated star-formation processes on a pc scale with astrophysical behavior of systems on galactic and cosmological scales.

14 The Universal Mass Function

How does the stellar IMF fit in with the mass distribution of all condensed objects, from planets to massive galaxy clusters?

Binggeli and Hascher (2007) deduce a universal MF (UMF) for all astronomical objects over a mass range of 36 decimal orders of magnitude, divided into seven groups: (1) planets and small bodies (meteoroids, asteroids); (2) stars, BDs, and stellar remnants; (3) molecular clouds; (4) open clusters; (5) globular clusters; (6) galaxies (dark matter plus baryonic masses); and (7) galaxy clusters (dark matter plus baryonic masses). Their result is reproduced in [Fig. 4-37](#). The shape as well as the normalization are based on observational data, however, using a present-day mass function rather than the IMF for stars, with a slightly different low-mass stellar slope and BDs as being continuously connected to stars (dotted curve). The canonical IMF is shown by the blue curves. The UMF follows an approximate $\propto m^{-2}$ power law which is remarkably similar to the Salpeter one ($\alpha = 2.35$) as well as to a logarithmic mass equi-distribution ($\alpha = 2$).

Binggeli and Hascher (2007) write “It is gratifying that the two halves almost perfectly connect to each other around one solar mass. Remember that this normalization was achieved on



■ Fig. 4-37

The universal mass function (UMF) of condensed structures according to Binggeli and Hascher (2007). BH07 estimated the absolute normalization based on the observed occurrence of the objects ranging from planets to galaxy clusters. The canonical IMF are the TK07 lines ([Fig. 4.55](#)), Thies and Kroupa 2007). Note that if dark matter does not exist, then the galaxy and galaxy cluster MFs are given by the *black line* (lower limit to the *shaded region*), while the *dark matter plus baryonic masses* of galaxies are given by the *upper envelope* of this *shaded region*

the basis of the mean universal mass density carried by the stars, embodied in the galaxies on the one hand, and comprised by the stars themselves on the other.” As pointed out by Binggeli and Hascher (2007), a possible reason for this fairly continuous UMF may be that gravitation is the dominant agent for creating the structures.

15 Concluding Comments

Spectacular advances have been achieved over the past two decades in the field of IMF research, and this affects a vast area of astrophysics. The discrepant observed nearby and distant luminosity functions of stars in the local Galactic disk have been unified with one IMF, and the strong peak in the luminosity function near $M_V = 12$ is well understood as a result of the changing internal constitution of stars as a function of their mass. The unification of the luminosity functions also ultimately lead to a unification of the observed discrepant binary populations in different environments through dynamical processing in star-forming regions.

The observationally determined stellar IMF is described by (☉ 4.55) and/or (☉ 4.56). Its form has traditionally been understood to be invariant in contradiction to the variable IMF prediction (p. 124). The IMF of stellar systems and of the Galactic field is provided by (☉ 4.57) and (☉ 4.59), respectively.

By studying the IMF, star-formation theory is being tested. According to the IMF Origin Conjecture (p. 207), the vast majority of stellar masses ($0.1 \lesssim m/M_\odot \lesssim \text{few}$) do not appear to be affected by competitive accretion or proto-stellar interactions (☉ Sect. 11.2). According to the computational star-formation research, the most massive star in a star-formation event correlates physically with the mass of the event. This $m_{\max} - M_{\text{ecl}}$ relation is a natural outcome of the competitive accretion-growth and the fragmentation-induced starvation of stellar masses, and the computations reproduce the general shape of the IMF. Different ideas (competitive accretion, coagulation and simply the distribution of gravitationally unstable regions in turbulent clouds) all lead to virtually the same type of theoretical power-law IMF. The IMF appears to be mostly form invariant within space-time correlated star-formation events (CSFEs, i.e., in individual embedded star clusters) for star-formation rate densities $SFRD \lesssim 0.1 M_\odot \text{ pc}^{-3} \text{ year}^{-1}$ within a spatial scale of about a pc. But, the IMF varies (trivially) among such individual CSFEs of stellar mass M_{ecl} through the $m_{\max} - M_{\text{ecl}}$ relation. This relation follows from the notion of optimal sampling. Computational star formation has achieved a remarkable degree of realism, although the binary population has not yet emerged properly and the computations cannot yet reach the birth of populous star clusters.

That optimal sampling appears to be describing a freshly born stellar population would invalidate the concept of a stellar population being a purely random representation of the IMF, leading to Open Question II (p. 150). The remarkable similarity of observationally determined IMF power-law indices (Open Question III, p. 195) may well be an indication that nature follows optimal sampling. Two other open questions related to star formation, and the IMF are furthermore stated (pp. 143 and 208).

Evidence for a variation of the shape of the IMF has emerged, being consistent with the long-previously predicted IMF variation. The evidence for top-heavy IMFs for $SFRD \gtrsim 0.1 M_\odot \text{ pc}^{-3} \text{ year}^{-1}$ (☉ 4.64) comes from either unresolved clusters or from populations that are very difficult to observe but appears to be increasingly established. The long-sought after evidence for a systematically varying stellar IMF with metallicity (☉ Fig. 4-33) may have emerged. A principal-component-type analysis of the PDMFs of GCs has now for the first

time yielded a formulation of a systematically varying IMF with star-forming cloud density and metallicity (► 4.65).

Among other intriguing recent results are that BDs appear to be a distinct population from that of low-mass stars; their pairing properties have a different energy scale. This is well reproduced by computational star formation. BDs and stars thus follow different mass distributions which do not join. A continuous log-normal function across the VLMS/BD mass scale does not therefore correctly describe the IMF.

Furthermore, the IMF does appear to have a physical maximum stellar mass that has now been found empirically. Stars with $m \gtrsim 150 M_{\odot}$ do not appear to form, unless they implode invisibly shortly after being formed. Populous star clusters with super-canonical ($m > 150 M_{\odot}$) stars are termed to be super saturated (p. 148). There is no statistically meaningful observational evidence for the formation of massive stars in isolation, in agreement with star-formation computations. The established invariance of the observationally derived IMF precludes the exotic IMF associated with isolated massive star formation.

By realizing that CSFEs are the true fundamental building blocks of a galaxy, all such events with their IMFs need to be added up to arrive at the integrated galactic initial mass function. This IGIMF varies in dependence of the SFR of the galaxy. According to the IGIMF theory, galaxies with low SFRs have a smaller ratio between the number of massive stars and low-mass stars than galaxies with high SFRs. This is the “IGIMF effect.” This formulation allows computation of the IGIMF as a function of time for galaxies with different SFRs. One implication of this is that equally old galaxies can have very different chemical compositions ranging from unevolved to evolved and that the cosmological supernova type II rate per star would be significantly different and dependent on galaxy type than if an invariant stellar IMF is assumed. Galaxy-formation and evolution computations with this latter assumption are not likely to be correct. Indeed, it transpires that only with the IGIMF theory is a fundamental time scale of about 3 Gyr uncovered on which all late-type galaxies consume their current gas supply. Very simple star-formation laws for galaxies emerge (► 4.73). Also, only with the IGIMF theory are the stellar-mass buildup times of dwarf galaxies consistent with the Hubble time. The top-heavy IMF of the Galactic Bulge deduced from chemical evolution research follows immediately from the IGIMF theory.

Thus, with the IGIMF theory, it has now become possible to calculate how the observationally well-constrained star formation on pc scales propagates through to galactic scales of cosmological relevance. A unification of scales has therewith been achieved which was quite unthinkable only a few years ago.

Many details still need to be worked out though. For example, direct verification of the IGIMF effect is needed. This can be achieved by directly counting the number of massive stars in dwarf galaxies with low SFRs or by counting the number of massive stars in a giant molecular cloud and comparing this number with the number of late-type stars formed there. Also, the existence of and exact form of the $m_{\max} - M_{\text{ecl}}$ relation is important not only for the calculation of the IGIMF but also for understanding to which degree star formation is self-regulated on a pc scale. Such work will establish which of the two philosophical approaches of ► Sect. 1.5 describe the astrophysics of star formation and of galaxies. The fact that the Orion southern cloud L1641 has formed a significant deficit of massive stars despite producing thousands of stars (Hsu et al. 2012) and the fact that the $m_{\max} - M_{\text{ecl}}$ relation is excellently mapped with a small scatter at the lowest masses (Kirk and Myers 2011) already constitute major evidence that the IGIMF theory and its fundamental assumptions appear to be given by nature.

Concerning the wider picture, the stellar and BD IMFs fit in quite continuously into the universal mass function of condensed structures as pointed out by Binggeli and Hascher (2007).

Acknowledgments

PK is thankful to Christopher Tout and Gerry Gilmore for very stimulating and important contributions without which much of this material would not have become available. PK is especially indebted to Sverre Aarseth whose friendly tutoring (against “payments” in the form of many bottles of *lieblichen* German white wine) eased the numerical dynamics work in 1993/1994. Douglas Heggie be thanked for fruitful discussions with PK on optimal sampling in Heidelberg in September, 2011. We thank Sambaran Banerjee for very helpful comments on the manuscript. PK would also like to thank M. R. S. Hawkins who had introduced him to this field in 1987 while PK visited the Siding-Spring Observatory as a summer vacation scholar at the ANU. Mike gave PK a delightful lecture on the low-mass LF one night when visiting his observing run to learn about modern, state-of-the-art Schmidt-telescope surveying with photographic plates *before* PK embarked on postgraduate work. This research was much later supported through DFG grants KR1635/2 and KR1635/3 and a Heisenberg fellowship, KR1635/4, KR1635/12, KR1635/13, and currently KR1635/25. MM acknowledges the Bonn/Cologne International Max-Planck Research School for support.

Cross-References

- [Binaries and Multiple Stellar Systems](#)
- [Brown Dwarfs](#)
- [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- [Dynamics of Disks and Warps](#)
- [Evolution of High-Mass Stars](#)
- [Evolution of Solar and Intermediate Mass Stars](#)
- [Galactic Distance Scales](#)
- [Globular Cluster Dynamical Evolution](#)
- [Mass Distribution and Rotation Curve in the Galaxy](#)
- [Metal-Poor Stars and the Chemical Enrichment of the Universe](#)
- [Numerical Techniques in Astrophysics](#)
- [Open Clusters and their Role in the Galaxy](#)
- [Star Counts and the Nature of the Galactic Thick Disk](#)
- [Star Formation](#)
- [Statistical Methods for Astronomy](#)
- [Stellar Populations](#)
- [The Galactic Bulge](#)
- [The Galactic Nucleus](#)
- [White Dwarf Stars](#)
- [Young Stellar Objects and Protostellar Disks](#)

References

- | | |
|---|--|
| Aarseth, S. J. 1999, <i>PASP</i> , 111, 1333 | Adams, F. C., & Fatuzzo, M. 1996, <i>ApJ</i> , 464, 256 |
| Aarseth, S. J. 2003, in <i>Gravitational N-Body Simulations</i> , ed. S. J. Aarseth (Cambridge, UK: Cambridge University Press), 430. ISBN 0521432723 | Adams, F. C., & Myers, P. C. 2001, <i>ApJ</i> , 553, 744 |
| | Allen, P. R., Koerner, D. W., Reid, I. N., & Trilling, D. E. 2005, <i>ApJ</i> , 625, 385 |

- Alves, J., Lombardi, M., & Lada, C. J. 2007, *A&A*, 462, L17
- Anders, P., Lamers, H. J. G. L. M., & Baumgardt, H. 2009, *A&A*, 502, 817
- Andersen, J. 1991, *A&A*, 3, 91
- Andersen, M., Meyer, M. R., Greissl, J., & Aversa, A. 2008, *ApJ*, 683, L183
- Andersen, M., Zinnecker, H., & Moneti, A., et al. 2009, *ApJ*, 707, 1347
- Anderson, J. P., Haberman, S. M., & James, P. A. 2011, *MNRAS*, 416, 567
- André, P., Belloche, A., Motte, F., & Peretto, N. 2007, *A&A*, 472, 519
- André, P., Men'shchikov, A., & Bontemps, S., et al. 2010, *A&A*, 518, L102+
- Bahcall, J. N. 1984, *ApJ*, 287, 926
- Ballero, S. K., Kroupa, P., & Matteucci, F. 2007a, *A&A*, 467, 117
- Ballero, S. K., Matteucci, F., Origlia, L., & Rich, R. M. 2007b, *A&A*, 467, 123
- Banerjee, S., & Kroupa, P. 2012, *A&A*, in press
- Banerjee, S., Kroupa, P., & Oh, S. 2012, *ApJ*, 746, 15
- Baraffe, I., & Chabrier, G. 2010, *A&A*, 521, A44+
- Baraffe, I., Chabrier, G., Allard, F., & Hauschildt, P. H. 1998, *A&A*, 337, 403
- Baraffe, I., Chabrier, G., Allard, F., & Hauschildt, P. H. 2002, *A&A*, 382, 563
- Baraffe, I., Chabrier, G., & Gallardo, J. 2009, *ApJ*, 702, L27
- Barrado y Navascués, D., Stauffer, J. R., Bouvier, J., & Martín, E. L. 2001, *ApJ*, 546, 1006
- Barrado y Navascués, D., Stauffer, J. R., & Jayawardhana, R. 2004, *ApJ*, 614, 386
- Bartko, H., Martins, F., & Trippel, S., et al. 2010, *ApJ*, 708, 834
- Basri, G. 2000, *ARA&A*, 38, 485
- Bastian, N., Covey, K. R., & Meyer, M. R. 2010, *ARA&A*, 48, 339
- Basu, S., & Jones, C. E. 2004, *MNRAS*, 347, L47
- Basu, S., & Vorobyov, E. I. 2012, *ApJ*, 750, 30
- Bate, M. R. 2005, *MNRAS*, 363, 363
- Bate, M. R. 2009, *MNRAS*, 397, 232
- Bate, M. R., & Bonnell, I. A. 2005, *MNRAS*, 356, 1201
- Baugh, C. M., Lacey, C. G., & Frenk, C. S., et al. 2005, *MNRAS*, 356, 1191
- Baumgardt, H., & Makino, J. 2003, *MNRAS*, 340, 227
- Baumgardt, H., De Marchi, G., & Kroupa, P. 2008, *ApJ*, 685, 247
- Beech, M., & Mitalas, R. 1994, *ApJS*, 95, 517
- Belikov, A. N., Hirte, S., Meusinger, H., Piskunov, A. E., & Schilbach, E. 1998, *A&A*, 332, 575
- Bestenlehner, J. M., Vink, J. S., & Gräfener, G., et al. 2011, *A&A*, 530, L14+
- Beuzit, J. L., Ségransan, D., & Forveille, T., et al. 2004, *A&A*, 425, 997
- Binggeli, B., & Hascher, T. 2007, *PASP*, 119, 592
- Binney, J., & Merrifield, M. 1998, in *Princeton Series in Astrophysics, Galactic Astronomy*, eds. J. Binney, & M. Merrifield (Princeton, NJ: Princeton University Press), QB857 .B522 1998 (\$35.00)
- Binney, J., Dehnen, W., & Bertelli, G. 2000, *MNRAS*, 318, 658
- Blumenthal, G. R., Faber, S. M., Flores, R., & Primack, J. R. 1986, *ApJ*, 301, 27
- Bochanski, J. J., Hawley, S. L., & Covey, K. R., et al. 2010, *AJ*, 139, 2679
- Boily, C. M., Lançon, A., Deiters, S., & Heggge, D. C. 2005, *ApJ*, 620, L27
- Boissier, S., Gil de Paz, A., & Boselli, A., et al. 2007, *ApJS*, 173, 524
- Bonnell, I. A., & Bate, M. R. 2002, *MNRAS*, 336, 659
- Bonnell, I. A., & Bate, M. R. 2006, *MNRAS*, 370, 488
- Bonnell, I. A., & Rice, W. K. M. 2008, *Science*, 321, 1060
- Bonnell, I. A., Bate, M. R., & Zinnecker, H. 1998, *MNRAS*, 298, 93
- Bonnell, I. A., Vine, S. G., & Bate, M. R. 2004, *MNRAS*, 349, 735
- Bonnell, I. A., Clarke, C. J., & Bate, M. R. 2006, *MNRAS*, 368, 1296
- Bonnell, I. A., Larson, R. B., & Zinnecker, H. 2007, *Protostars and Planets V*, 149
- Bontemps, S., André, P., Kaas, A. A. 2001, *A&A*, 372, 173
- Bontemps, S., Motte, F., Csengeri, T., & Schneider, N. 2010, *A&A*, 524, A18+
- Bosch, G., Selman, F., Melnick, J., & Terlevich, R. 2001, *A&A*, 380, 137
- Boss, A. R. 1986, *ApJS*, 62, 519
- Bouvier, J., Moraux, E., Stauffer, J. R., Barrado y Navascués, D., & Cuillandre, J. C. 2003, in *IAU Symposium*, 147, *Brown Dwarfs*, vol. 211 ed. E. Martín, astro-ph/0209178
- Bouvier, J., Kendall, T., & Meeus, G., et al. 2008, *A&A*, 481, 661
- Bouy, H., Brandner, W., & Martín, E. L., et al. 2003, *AJ*, 126, 1526
- Briceño, C., Luhman, K. L., Hartmann, L., Stauffer, J. R., & Kirkpatrick, J. D. 2002, *ApJ*, 580, 317
- Brocato, E., Cassisi, S., & Castellani, V. 1998, *MNRAS*, 295, 711
- Bromm, V., Ferrara, A., Coppi, P. S., Larson, R. B. 2001, *MNRAS*, 328, 969
- Brüns, R. C., Kroupa, P., Fellhauer, M., Metz, M., & Assmann, P. 2011, *A&A*, 529, A138+
- Calzetti, D., Chandar, R., & Lee, J. C., et al. 2010, *ApJ*, 719, L158
- Camargo, D., Bonatto, C., & Bica, E. 2010, *A&A*, 521, A42+

- Casuso, E., & Beckman, J. E. 2012, *MNRAS*, 419, 1642
- Cenarro, A. J., Gorgas, J., Vazdekis, A., Cardiel, N., & Peletier, R. F. 2003, *MNRAS*, 339, L12
- Chabrier, G. 2001, *ApJ*, 554, 1274
- Chabrier, G. 2002, *ApJ*, 567, 304
- Chabrier, G. 2003a, *ApJ*, 586, L133
- Chabrier, G. 2003b, *PASP*, 115, 763
- Chabrier, G., & Baraffe, I. 1997, *A&A*, 327, 1039
- Chabrier, G., & Baraffe, I. 2000, *ARA&A*, 38, 337
- Chini, R., Hoffmeister, V., & Kimeswenger, S., et al. 2004, *Nature*, 429, 155
- Chini, R., Hoffmeister, V. H., Nasserri, A., Stahl, O., & Zinnecker, H. 2012, *MNRAS*, in press, arXiv:astro-ph/1205.5238
- Chlebowski, T., & Garmany, C. D. 1991, *ApJ*, 368, 241
- Clarke, C. J., & Pringle, J. E. 1992, *MNRAS*, 255, 423
- Clark, P. C., Glover, S. C. O., Bonnell, I. A., & Klessen, R. S. 2009, *ApJ*, submitted, arXiv:astro-ph/0904.3302
- Clark, P. C., Glover, S. C. O., Klessen, R. S., & Bromm, V. 2011, *ApJ*, 727, 110
- Close, L. M., Siegler, N., Freed, M., & Biller, B. 2003, *ApJ*, 587, 407
- Conroy, C. 2011, *ApJ*, submitted, arXiv:astro-ph/1101.2208
- Corbelli, E., Palla, F., Zinnecker, H. (eds.), 2005, *The Initial Mass Function 50 years later, Astrophysics and Space Science Library*, 327 (Dordrecht: Springer)
- Corbelli, E., Verley, S., Elmegreen, B. G., & Giovannardi, C. 2009, *A&A*, 495, 479
- Crowther, P. A., Schnurr, O., & Hirschi, R., et al. 2010, *MNRAS*, 408, 731
- Dabringhausen, J., Kroupa, P., & Baumgardt, H. 2009, *MNRAS*, 394, 1529
- Dabringhausen, J., Fellhauer, M., & Kroupa, P. 2010, *MNRAS*, 403, 1054
- Dabringhausen, J., Kroupa, P., Pflamm-Altenburg, J., & Mieske, S. 2012, *ApJ*, 747, 72
- Dale, J. E., Wunsch, R., Smith, R. J., Whitworth, A., & Palouš, J., 2011, *MNRAS*, 411, 2230
- D'Antona, F., & Mazzitelli, I. 1996, *ApJ*, 456, 329
- Da Rio, N., Gouliermis, D. A., Henning, T. 2009, *ApJ*, 696, 528
- de Boer, K. S., Fitzpatrick, E. L., & Savage, B. D. 1985, *MNRAS*, 217, 115
- de La Fuente Marcos, R. 1998, *A&A*, 333, L27
- de Marchi, G., & Paresce, F. 1995a, *A&A*, 304, 202
- de Marchi, G., & Paresce, F. 1995b, *A&A*, 304, 211
- De Marchi, G., Paresce, F., & Pulone, L. 2007, *ApJ*, 656, L65
- de Wit, W. J., Testi, L., Palla, F., Vanzì, L., & Zinnecker, H. 2004, *A&A*, 425, 937
- de Wit, W. J., Testi, L., Palla, F., & Zinnecker, H. 2005, *A&A*, 437, 247
- Deason, A. J., Belokurov, V., Evans, N. W., & McCarthy, I. G. 2011, *ApJ*, in press, arXiv:astro-ph/1110.0833
- Delfosse, X., Forveille, T., & Beuzit, J. L., et al. 1999, *A&A*, 344, 897
- Delfosse, X., Forveille, T., & Ségransan, D., et al. 2000, *A&A*, 364, 217
- Dib, S. 2011, *ApJ*, 737, L20+
- Dib, S., Shadmehri, M., & Padoan, P., et al. 2010, *MNRAS*, 405, 401
- Dib, S., Piau, L., Mohanty, S., & Braine, J. 2011, *MNRAS*, 415, 3439
- Disney, M. J., Romano, J. D., & Garcia-Appadoo, D. A., et al. 2008, *Nature*, 455, 1082
- Djorgovski, S., Piotto, G., & Capaccioli, M. 1993, *AJ*, 105, 2148
- Dominik, M. 2011, *MNRAS*, 411, 2
- Duchêne, G., Simon, T., Eisloffel, J., & Bouvier, J. 2001, *A&A*, 379, 147
- Duquenois, A., & Mayor, M. 1991, *A&A*, 248, 485
- Eddington, A. S., 1926, *The Internal Constitution of the Stars* (Cambridge: Cambridge University Press)
- Egusa, F., Sofue, Y., & Nakanishi, H. 2004, *PASJ*, 56, L45
- Eisenhauer, F. 2001, in *Science with the Large Binocular Telescope*, eds. T. Herbst, 89, arXiv:astro-ph/0101384
- Elmegreen, B. G. 1983, *MNRAS*, 203, 1011
- Elmegreen, B. G. 1997, *ApJ*, 486, 944
- Elmegreen, B. G. 1999, *ApJ*, 515, 323
- Elmegreen, B. G. 2000, *ApJ*, 539, 342
- Elmegreen, B. G. 2004, *MNRAS*, 354, 367
- Elmegreen, B. G. 2005, in *Astrophysics and Space Science Library*, 57, *Starbursts: From 30 Doradus to Lyman Break Galaxies*, Vol. 329, eds. R. de Grijs & R. M. González Delgado (Dordrecht/New York: Springer), arXiv:astro-ph/0411193
- Elmegreen, B. G. 2009, in *The Evolving ISM in the Milky Way and Nearby Galaxies*, arXiv:astro-ph/0803.3154
- Elmegreen, B. G. 2011, *ApJ*, 731, 61
- Elmegreen, B. G., & Scalzo, J. 2006, *ApJ*, 636, 149
- Elmegreen, B. G., & Shadmehri, M. 2003, *MNRAS*, 338, 817
- Elmegreen, B. G., Klessen, R. S., & Wilson, C. D. 2008, *ApJ*, 681, 365
- Esteban, C., Peimbert, M., Torres-Peimbert, S., & Escalante, V. 1998, *MNRAS*, 295, 401
- Evstigneeva, E. A., Gregg, M. D., Drinkwater, M. J., & Hilker, M. 2007, *AJ*, 133, 1722
- Fellhauer, M., & Kroupa, P. 2002, *MNRAS*, 330, 642
- Fellhauer, M., & Kroupa, P. 2005, *MNRAS*, 359, 223
- Fellhauer, M., Lin, D. N. C., Bolte, M., Aarseth, S. J., & Williams, K. A. 2003, *ApJ*, 595, L53

- Fellhauer, M., Wilkinson, M. I., & Kroupa, P. 2009, *MNRAS*, 397, 954
- Feltzing, S., Gilmore, G., & Wyse, R. F. G. 1999, *ApJ*, 516, L17
- Figer, D. F. 2003, in *IAU Symposium*, 487, A Massive Star Odyssey: From Main Sequence to Supernova, eds. K. van der Hucht, A. Herrero, & C. Esteban, Vol. 212 (San Francisco, CA: ASP)
- Figer, D. F. 2005, *Nature*, 434, 192
- Finkelstein, K. D., Papovich, C., & Finkelstein, S. L., et al. 2011, *ApJ*, 742, 108
- Fischer, D. A., & Marcy, G. W. 1992, *ApJ*, 396, 178
- Flynn, C., & Fuchs, B. 1994, *MNRAS*, 270, 471
- Forbes, D. A., & Kroupa, P. 2011, *PASA*, 28, 77
- Fuhrmann, K. 2004, *Astron Nachr*, 325, 3
- Fumagalli, M., da Silva, R. L., & Krumholz, M. R. 2011, *ApJ*, 741, L26
- García, B., & Mermillod, J. C. 2001, *A&A*, 368, 122
- García-Segura, G., Langer, N., & Mac Low, M. M. 1996a, *A&A*, 316, 133
- García-Segura, G., Mac Low, M. M., & Langer, N. 1996b, *A&A*, 305, 229
- Gilmore, G., Howell, D. (eds.), 1998, *The Stellar Initial Mass Function (38th Herstmonceux Conference)* (San Francisco, CA: ASP) ASP Conference Series, Vol. 142
- Gilmore, G. F., Perryman, M. A., & Lindegren, L., et al. 1998, in *Proc. SPIE*, 3350, *Astronomical Interferometry*, ed. R. D. Reasenberg (Bellingham, WA: SPIE), 541–550, arXiv:astro-ph/9805180
- Gizis, J. E., Kirkpatrick, J. D., & Burgasser, A., et al. 2001, *ApJ*, 551, L163
- Girichidis, P., Federrath, C., Banerjee, R., & Klessen, R. S. 2011, *MNRAS*, 413, 2741
- Goodwin, S. P., & Bastian, N. 2006, *MNRAS*, 373, 752
- Goodwin, S. P., & Kroupa, P. 2005, *A&A*, 439, 565
- Goodwin, S. P., & Pagel, B. E. J. 2005, *MNRAS*, 359, 707
- Goodwin, S. P., & Whitworth, A. 2007, *A&A*, 466, 943
- Goodwin, S. P., Kroupa, P., Goodman, A., & Burkert, A. 2007, in *Protostars and Planets V*, eds. B. Reipurth, D. Jewitt, & K. Keil (Tucson: University of Arizona Press), 133–147, arXiv:astro-ph/0603233
- Goodwin, S. P., Nutter, D., Kroupa, P., Ward-Thompson, D., & Whitworth A. P. 2008, *A&A*, 477, 823
- Gouliermis, D., Brandner, W., & Henning T. 2005, *ApJ*, 623, 846
- Gouliermis, D., Brandner, W., & Henning, T. 2006, *ApJ*, 636, L133
- Gouliermis, D. A., Bestenlehner, J. M., Brandner, W., & Henning, T. 2010, *A&A*, 515, A56+
- Grillmair, C. J., Mould, J. R., & Holtzman, J. A. 1998, *AJ*, 115, 144
- Grillo, C., & Gobat, R. 2010, *MNRAS*, 402, L67
- Gvaramadze, V. V., & Bomans, D. J. 2008, *A&A*, 490, 1071
- Gvaramadze, V. V., & Gualandris, A. 2011, *MNRAS*, 410, 304
- Gvaramadze, V. V., Weidner, C., Kroupa, P., & Pflamm-Altenburg, J. 2012, *MNRAS*, in press, arXiv:astro-ph/1206.1596
- Haas, M. R., & Anders, P. 2010, *A&A*, 512, A79+
- Halbwachs, J. L., Arenou, F., Mayor, M., Udry, S., & Queloz, D. 2000, *A&A*, 355, 581
- Hambly, N. C., Jameson, R. F., & Hawkins, M. R. S. 1991, *MNRAS*, 253, 1
- Hambly, N. C., Hodgkin, S. T., Cossburn, M. R., & Jameson, R. F. 1999, *MNRAS*, 303, 835
- Hayashi, C., & Nakano, T. 1963, *Prog Theor Phys*, 30, 460
- Haywood, M., Robin, A. C., & Creze, M. 1997, *A&A*, 320, 440
- Heggie, D. C. 1975, *MNRAS*, 173, 729
- Hennebelle, P., & Chabrier, G. 2008, *ApJ*, 684, 395
- Hennebelle, P., & Chabrier, G. 2009, *ApJ*, 702, 1428
- Henry, T. J., Ianna, P. A., Kirkpatrick, J. D., & Jahreiss, H. 1997, *AJ*, 114, 388
- Hillenbrand, L. A. 1997, *AJ*, 113, 1733
- Hillenbrand, L. A. 2004, in *The Dense Interstellar Medium in Galaxies*, eds. S. Pfalzner et al. (Berlin/New York: Springer), 601, arXiv:astro-ph/0312187
- Hillenbrand, L. A., & Carpenter, J. M. 2000, *ApJ*, 540, 236
- Hocuk, S., & Spaans, M. 2011, *A&A*, 536, A41
- Hollenbach, D., Parravano, A., & McKee, C. F. 2005, in *ASSL*, 327, *The Initial Mass Function 50 Years Later*, eds. E. Corbelli, F. Palla, & H. Zinnecker (New York: Springer), 417–424
- Hoversten, E. A., & Glazebrook, K. 2008, *ApJ*, 675, 163
- Hsu, W.-H., Hartmann, L., Allen, L., et al. 2012, *ApJ*, 752, 59
- Ivanova, N., Belczynski, K., Fregeau, J. M., & Rasio, F. A. 2005, *MNRAS*, 358, 572
- Jahreiss, H. 1994, *Ap&SS*, 217, 63
- Jahreiß, H., & Wielen, R. 1997, in *ESA SP-402: Hipparcos – Venice '97*, ed. by B. Battarick et al. (Noordwijk: ESA Publications Division), 675–680
- Janka, H. T. 2001, *A&A*, 368, 527
- Jao, W., Henry, T. J., & Subasavage, J. P., et al. 2003, *AJ*, 125, 332
- Jeans, J. H. 1902, *R Soc Lond Philos Trans Ser A*, 199, 1
- Jijina, J., & Adams, F. C. 1996, *ApJ*, 462, 874

- Joergens, V. 2008, *A&A*, 492, 545
- Kahn, F. D. 1974, *A&A*, 37, 149
- Kalirai, J. S., Hansen, B. M. S., & Kelson, D. D. 2008, *ApJ*, 676, 594
- Kaplan, M., Stamatellos, D., & Whitworth, A. P. 2012, *Ap&SS*, 222, in press, arXiv:astro-ph/1205.2279
- Kennicutt, R. C., Jr. 2008, in *ASP Conf. Ser.* 149, *Pathways Through an Eclectic Universe*, eds. J. H. Knapen, T. J. Mahoney, & A. Vazdekis, Vol. 390 (San Francisco, CA: ASP)
- Kennicutt, R. C., Jr., Tamblyn, P., & Congdon, C. E. 1994, *ApJ*, 435, 22
- Kevlahan, N., & Pudritz, R. E. 2009, *ApJ*, 702, 39
- Kim, S. S., Figer, D. F., Kudritzki, R. P., & Najarro, F. 2006, *ApJ*, 653, L113
- Kirk, H., & Myers, P. C. 2011, *ApJ*, 727, 64
- Kirk, J. M., Ward-Thompson, D., & André, P. 2005, *MNRAS*, 360, 1506
- Klessen, R. S. 2001, *ApJ*, 550, L77
- Klessen, R. S., Spaans, M., & Jappsen, A. 2007, *MNRAS*, 374, L29
- Koen, C. 2006, *MNRAS*, 365, 590
- Köppen, J., Weidner, C., & Kroupa, P. 2007, *MNRAS*, 375, 673
- Kroupa, P. 1995a, *ApJ*, 453, 350
- Kroupa, P. 1995b, *MNRAS*, 277, 1522
- Kroupa, P. 1995c, *MNRAS*, 277, 1507
- Kroupa, P. 1995d, *MNRAS*, 277, 1491
- Kroupa, P. 1995e, *ApJ*, 453, 358
- Kroupa, P. 2000, *New Astron.*, 4, 615
- Kroupa, P. 2001a, *MNRAS*, 322, 231
- Kroupa, P. 2001b, in *ASP Conf. Ser.* 228, *Dynamics of Star Clusters and the Milky Way*, eds. S. Deiters et al. (San Francisco, CA: ASP), 187, arXiv:astro-ph/0011328
- Kroupa, P. 2001c, in: *IAU Symposium* 200, 199, arXiv:astro-ph/0010347
- Kroupa, P. 2002, *Science*, 295, 82
- Kroupa P., Jan. 2005, in *Proceedings of the Gaia Symposium "The Three-Dimensional Universe with Gaia"* (ESA SP-576). Held at the Observatoire de Paris-Meudon, 4–7 October 2004, eds. C. Turon, K. S. O’Flaherty, & M. A. C. Perryman (Noordwijk: ESA Publications Division), 629, arXiv:astro-ph/0412069
- Kroupa, P. 2008, in *LNP* 760, *The Cambridge N-Body Lectures*, eds. S. J. Aarseth, C. A. Tout, & R. A. Mardling (Berlin, Springer), 181, arXiv:astro-ph/0803.1833
- Kroupa, P., Famaey, B., & de Boer, K. S., et al. 2010, *A&A*, 523, A32+
- Kroupa, P. 2011, in *IAU Symposium*, Vol. 270, eds. J. Alves, B. G. Elmegreen, J. M. Girart, & V. Trimble, 141–149, arXiv:astro-ph/1012.1596
- Kroupa, P. 2012, *PASA*, in press, arXiv:astro-ph/1204.2546
- Kroupa, P., & Bouvier, J. 2003a, *MNRAS*, 346, 343
- Kroupa, P., & Bouvier, J. 2003b, *MNRAS*, 346, 369
- Kroupa, P., & Tout, C. A. 1997, *MNRAS*, 287, 402
- Kroupa, P., & Weidner, C. 2003, *ApJ*, 598, 1076
- Kroupa, P., Tout, C. A., & Gilmore G. 1990, *MNRAS*, 244, 76
- Kroupa, P., Gilmore, G., & Tout, C. A. 1991, *MNRAS*, 251, 293
- Kroupa, P., Tout, C. A., & Gilmore, G. 1993, *MNRAS*, 262, 545
- Kroupa, P., Aarseth, S., & Hurley, J. 2001, *MNRAS*, 321, 699
- Kroupa, P., Bouvier, J., Duchêne, G., & Moraux, E. 2003, *MNRAS*, 346, 354
- Krumholz, M. R., & McKee, C. F. 2008, *Nature*, 451, 1082
- Krumholz, M. R., Klein, R. I., McKee, C. F., Offner, S. S. R., & Cunningham A. J. 2009, *Science*, 323, 754
- Krumholz, M. R., Cunningham, A. J., Klein, R. I., & McKee, C. F. 2010, *ApJ*, 713, 1120
- Kudritzki, R., & Puls, J. 2000, *ARA&A*, 38, 613
- Kuijken, K., & Gilmore, G. 1991, *ApJ*, 367, L9
- Kumar, S. S. 2003, in *IAUS* 211, 3, arXiv:astro-ph/0208096
- Lada, C. J., & Lada, E. A. 2003, *ARA&A*, 41, 57
- Lamb, J. B., Oey, M. S., Werk, J. K., & Ingleby, L. D. 2010, *ApJ*, 725, 1886
- Larson, R. B. 1982, *MNRAS*, 200, 159
- Larson, R. B. 1998, *MNRAS*, 301, 569
- Larson, R. B. 2003, in *ASP Conf. Ser.* 287, *Galactic Star Formation Across the Stellar Mass Spectrum*, eds. J. M. De Buizer, & N. S. van der Bliek (San Francisco, CA: ASP), 65–80
- Lee, J. C., Gil de Paz, A., & Tremonti, C., et al. 2009, *ApJ*, 706, 599
- Li, Y., Klessen, R. S., & Mac Low, M. M. 2003, *ApJ*, 592, 975
- Li, Z.-Y., Wang, P., Abel, T., & Nakamura, F. 2010, *ApJ*, 720, L26
- Löckmann, U., Baumgardt, H., & Kroupa, P. 2010, *MNRAS*, 402, 519
- Lodieu, N., Dobbie, P. D., Deacon, N. R., et al. 2007, *MNRAS*, 380, 712
- Lonsdale, C. J., Diamond, P. J., Thrall, H., Smith, H. E., & Lonsdale, C. J. 2006, *ApJ*, 647, 185
- Low, C., & Lynden-Bell, D. 1976, *MNRAS*, 176, 367
- Luhman, K. L. 2004, *ApJ*, 617, 1216
- Luhman, K. L., Rieke, G. H., & Young, E. T., et al. 2000, *ApJ*, 540, 1016
- Lutz, T. E., & Kelker, D. H. 1973, *PASP*, 85, 573
- Machida, M. N., & Matsumoto, T. 2011, *MNRAS*, 421, 588

- Machida, M. N., Omukai, K., Matsumoto, T., & Inutsuka, S. 2009, *MNRAS*, 399, 1255
- Mac Low, M., & Klessen, R. S. 2004, *Rev Mod Phys*, 76, 125
- Maeder, A., & Behrend, R. 2002, in *ASP Conf. Ser. 267, Hot Star Workshop III: The Earliest Phases of Massive Star Birth*, ed. P. A. Crowther (San Francisco, CA: ASP), 179
- Maeder, A., & Meynet, G. 2000, *ARA&A*, 38, 143
- Maíz Apellániz, J. 2008, *ApJ*, 677, 1278
- Maíz Apellániz, J., & Úbeda, L. 2005, *ApJ*, 629, 873
- Maíz Apellániz, J., Walborn, N. R., Morrell, N. I., Niemela, V. S., & Nelan, E. P. 2007, *ApJ*, 660, 1480
- Maíz Apellániz, J., Walborn, N. R., Morrell, N. I., et al. 2008, in *Revista Mexicana de Astronomía y Astrofísica Conference Series* 33, 55–55 arXiv:astro-ph/0702514
- Malkov, O., & Zinnecker, H. 2001, *MNRAS*, 321, 149
- Malkov, O. Y., Piskunov, A. E., & Shpil’Kina, D. A. 1997, *A&A*, 320, 79
- Marks, M., & Kroupa, P. 2010, *MNRAS*, 406, 2000
- Marks, M., & Kroupa, P. 2011, *MNRAS*, 417, 1702
- Marks, M., & Kroupa, P. 2012, *A&A*, 543, A8
- Marks, M., Kroupa, P., & Baumgardt, H. 2008, *MNRAS*, 386, 2047
- Marks, M., Kroupa, P., & Oh, S. 2011, *MNRAS*, 417, 1684
- Marks, M., Kroupa, P., Dabringhausen, J., & Pawlowski, M. S. 2012, *MNRAS*, 422, 2246
- Martín, E. L., Barrado y Navascués, D., Baraffe, I., Bouy, H., & Dahm, S. 2003, *ApJ*, 594, 525
- Maschberger, T., & Clarke, C. J. 2008, *MNRAS*, 391, 711
- Maschberger, T., & Clarke, C. J. 2011, *MNRAS*, 1177
- Maschberger, T., & Kroupa, P. 2009, *MNRAS*, 395, 931
- Maschberger, T., Clarke, C. J., Bonnell, I. A., & Kroupa, P. 2010, *MNRAS*, 404, 1061
- Massey, P. 1998, in *ASP Conf. Ser. 142, The Stellar Initial Mass Function (38th Herstmonceux Conference)*, eds. G. Gilmore, & D. Howell (San Francisco, CA: ASP), 17
- Massey, P. 2003, *ARA&A*, 41, 15
- Massey, P., & Hunter, D. A. 1998, *ApJ*, 493, 180
- Maxted, P. F. L., & Jeffries, R. D. 2005, *MNRAS*, 362, L45
- Mayor, M., Duquennoy, A., Halbwachs, J. L., & Mermilliod, J. C. 1992, in *ASP Conf. Ser. 32, IAU Colloq. 135: Complementary Approaches to Double and Multiple Star Research*, eds. H. A. McAlister, & W. I. Hartkopf (San Francisco, CA: ASP), 73
- McCraday, N., Graham, J. R., & Vacca, W. D. 2005, *ApJ*, 621, 278
- McMillan, S. L. W., Vesperini, E., & Portegies Zwart, S. F. 2007, *ApJ*, 655, L45
- Megeath, S. T., Herter, T., & Beichman, C., et al. 1996, *A&A*, 307, 775
- Meurer, G. R., Wong, O. I., & Kim, J. H., et al. 2009, *ApJ*, 695, 765
- Meyer, M. R., Adams, F. C., Hillenbrand, L. A., Carpenter, J. M., & Larson, R. B. 2000, in *Protostars and Planets IV*, eds. V. Mannings, A. Boss, & S. S. Russell (Tucson: University of Arizona Press), 121
- Mieske, S., & Kroupa, P. 2008, *ApJ*, 677, 276
- Mieske, S., Hilker, M., & Infante, L. 2002, *A&A*, 383, 823
- Miller, G. E., & Scalo, J. M. 1979, *ApJS*, 41, 513
- Misgeld, I., & Hilker, M. 2011, *MNRAS*, 414, 3699
- Moeckel, N., & Bate, M. R. 2010, *MNRAS*, 404, 721
- Morau, E., Kroupa, P., & Bouvier, J. 2004, *A&A*, 426, 75
- Motte, F., Andre, P., & Neri, R. 1998, *A&A*, 336, 150
- Muench, A. A., Lada, E. A., & Lada, C. J. 2000, *ApJ*, 533, 358
- Murray, N. 2009, *ApJ*, 691, 946
- Murray, S. D., & Lin, D. N. C. 1996, *ApJ*, 467, 728
- Myers, P. C. 2011, *ApJ*, 743, 98
- Myers, A. T., Krumholz, M. R., Klein, R. I., & McKee, C. F. 2011, *ApJ*, 735, 49
- Nagashima, M., Lacey, C. G., Baugh, C. M., Frenk, C. S., & Cole, S. 2005, *MNRAS*, 358, 1247
- Najarro, F., Figer, D. F., Hillier, D. J., & Kudritzki, R. P. 2004, *ApJ*, 611, L105
- Nayakshin, S., & Sunyaev, R. 2005, *MNRAS*, 364, L23
- Neill, J. D., Sullivan, M., & Gal-Yam, A., et al. 2011, *ApJ*, 727, 15
- Nordlund, Å., & Padoan, P. 2003, *LNP* 614, *Turbulence and Magnetic Fields in Astrophysics*, eds. E. Falgarone & T. Passot (Berlin/New York: Springer), 271
- Oey, M. S. 2011, *ApJ*, 739, L46
- Oey, M. S., & Clarke, C. J. 2005, *ApJ*, 620, L43
- Oliveira, J. M., Jeffries, R. D., & van Loon, J. T. 2009, *MNRAS*, 392, 1034
- Onishi, T., Mizuno, A., Kawamura, A., Tachihara, K., & Fukui Y. 2002, *ApJ*, 575, 950
- Padoan, P., & Nordlund, Å. 2002, *ApJ*, 576, 870
- Padoan, P., & Nordlund, Å. 2004, *ApJ*, 617, 559
- Palla, F., Randich, S., Pavlenko, Y. V., Flaccomio, E., & Pallavicini, R. 2007, *ApJ*, 659, L41
- Papadopoulos, P. P. 2010, *ApJ*, 720, 226
- Papadopoulos, P. P., Thi, W. F., Miniati, F., & Viti, S. 2011, *MNRAS*, 414, 1705
- Paresce, F., de Marchi, G., & Romaniello, M. 1995, *ApJ*, 440, 216
- Parker, R. J., & Goodwin, S. P. 2007, *MNRAS*, 380, 1271
- Parker, R. J., & Goodwin, S. P. 2011, *MNRAS*, 411, 891

- Parker, R. J., Bouvier, J., & Goodwin, S. P., et al. 2011, *MNRAS*, 412, 2489
- Parra, R., Conway, J. E., & Diamond, P. J., et al. 2007, *ApJ*, 659, 314
- Parravano, A., McKee, C. F., & Hollenbach, D. J. 2011, *ApJ*, 726, 27
- Peebles, P. J. E., & Nusser, A. 2010, *Nature*, 465, 565
- Penny, L. R., Massey, P., & Vukovich, J. 2001, *Bull Am Astron Soc*, 33, 1310
- Pérez-Torres, M. A., Romero-Cañizales, C., Alberdi, A., & Polatidis, A. 2009, *A&A*, 507, L17
- Peters, T., Klessen, R. S., Mac Low, M. M., & Banerjee, R. 2010, *ApJ*, 725, 134
- Peters, T., Banerjee, R., Klessen, R. S., & Mac Low, M. M. 2011a, *ApJ*, 729, 72
- Peters, T., Klessen, R. S., Mac Low, M. M., & Banerjee, R. 2011b, *arXiv:astro-ph/1110.2892*
- Pflamm-Altenburg, J., & Kroupa, P. 2006, *MNRAS*, 373, 295
- Pflamm-Altenburg, J., & Kroupa, P. 2007, *MNRAS*, 375, 855
- Pflamm-Altenburg, J., & Kroupa, P. 2008, *Nature*, 455, 641
- Pflamm-Altenburg, J., & Kroupa, P. 2009a, *MNRAS*, 397, 488
- Pflamm-Altenburg, J., & Kroupa, P. 2009b, *ApJ*, 706, 516
- Pflamm-Altenburg, J., & Kroupa, P. 2010, *MNRAS*, 404, 1564
- Pflamm-Altenburg, J., Weidner, C., & Kroupa, P. 2007, *ApJ*, 671, 1550
- Pflamm-Altenburg, J., Weidner, C., & Kroupa, P. 2009, *MNRAS*, 395, 394
- Phan-Bao, N., Martin, E. L., Reylé, C., Forveille, T., & Lim, J. 2005, *A&A*, 439, L19
- Pinfield, D. J., Dobbie, P. D., & Jameson, R. F., et al. 2003, *MNRAS*, 342, 1241
- Piotto, G., & Zoccali, M. 1999, *A&A*, 345, 485
- Piskunov, A. E., Belikov, A. N., Kharchenko, N. V., Sagar, R., & Subramaniam, A. 2004, *MNRAS*, 349, 1449
- Portegies Zwart, S. F., Makino, J., McMillan, S. L. W., & Hut, P. 2002, *ApJ*, 565, 265
- Preibisch, T., Balega, Y., Hofmann, K., Weigelt, G., & Zinnecker, H. 1999, *New Astron*, 4, 531
- Price, D. J., & Bate, M. R. 2009, *MNRAS*, 398, 33
- Ramspeck, M., Heber, U., & Moehler, S. 2001, *A&A*, 378, 907
- Recchi, S., Calura, F., & Kroupa, P. 2009, *A&A*, 499, 711
- Reid, I. N., & Gizis, J. E. 1997, *AJ*, 113, 2246
- Reid, I. N., Gizis, J. E., & Hawley, S. L. 2002, *AJ*, 124, 2721
- Reid, I. N., Cruz, K. L., & Allen, P., et al. 2003a, *AJ*, 126, 3007
- Reid, I. N., Cruz, K. L., & Laurie, S. P., et al. 2003b, *AJ*, 125, 354
- Reid, N., & Gilmore, G. 1982, *MNRAS*, 201, 73
- Reipurth, B., & Clarke, C. 2001, *AJ*, 122, 432
- Reylé, C., & Robin, A. C. 2001, *A&A*, 373, 886
- Röser, S., Schilbach, E., Piskunov, A. E., Kharchenko, N. V., & Scholz, R. D. 2011, *A&A*, 531, A92
- Roychowdhury, S., Chengalur, J. N., Kaisin, S. S., Begum, A., & Karachentsev, I. D. 2011, *MNRAS*, 414, L55
- Salpeter, E. E. 1955, *ApJ*, 121, 161
- Sana, H., & Evans, C. J. 2010, *arXiv:astro-ph/1009.4197*
- Sana, H., James, G., & Gosset, E. 2011a, *MNRAS*, 416, 817
- Sana, H., Lacour, S., & Le Bouquin, J., et al. 2011b, *arXiv:astro-ph/1109.6654*
- Scalo, J. M. 1986, *Fundam Cosm Phys*, 11, 1
- Scalo, J. 1998, in *ASP Conf. Ser. 142, The Stellar Initial Mass Function (38th Herstmonceux Conference)*, eds. G. Gilmore & D. Howell (San Francisco, CA: ASP), 201
- Schilbach, E., & Röser, S. 2008, *A&A*, 489, 105
- Schmalzl, M., Gouliermis, D. A., Dolphin, A. E., & Henning, T. 2008, *ApJ*, 681, 290
- Schwarzschild, M., & Härm, R. 1959, *ApJ*, 129, 637
- Selier, R., Heydari-Malayeri, M., & Gouliermis, D. A. 2011, *A&A*, 529, A40+
- Selman, F., Melnick, J., Bosch, G., & Terlevich, R. 1999, *A&A*, 347, 532
- Shen, S., Wadsley, J., Hayfield, T., & Ellens, N. 2010, *MNRAS*, 401, 727
- Siess, L., Dufour, E., & Forestini, M. 2000, *A&A*, 358, 593
- Sirianni, M., Nota, A., Leitherer, C., De Marchi, G., & Clampin, M. 2000, *ApJ*, 533, 203
- Slesnick, C. L., Hillenbrand, L. A., & Carpenter, J. M. 2004, *ApJ*, 610, 1045
- Smith, L. J., & Gallagher, J. S. 2001, *MNRAS*, 326, 1027
- Smith, R. J., Longmore, S., & Bonnell, I. 2009, *MNRAS*, 400, 1775
- Soubiran, C., Bienaymé, O., & Siebert, A. 2003, *A&A*, 398, 141
- Sollima, A., Beccari, G., Ferraro, F. R., Fusi Pecci, F., & Sarajedini, A. 2007, *MNRAS*, 380, 781
- Sollima, A., Carballo-Bello, J. A., Beccari, G., et al. 2010, *MNRAS*, 401, 577
- Stahler, S. W., Palla, F., & Ho, P. T. P. 2000, *Protostars and Planets IV*, eds. V. Mannings, A. Boss, & S. S. Russell (Tucson: University of Arizona Press), 327

- Stamatellos, D., & Whitworth, A. P. 2009, *MNRAS*, 392, 413
- Stamatellos, D., Hubber, D. A., & Whitworth, A. P. 2007a, *MNRAS*, 382, L30
- Stamatellos, D., Whitworth, A. P., Bisbas, T., & Goodwin, S. 2007b, *A&A*, 475, 37
- Stamatellos, D., Whitworth, A. P., & Hubber, D. A. 2011, *ApJ*, 730, 32
- Stobie, R. S., Ishida, K., & Peacock, J. A. 1989, *MNRAS*, 238, 709
- Stolte, A., Grebel, E. K., Brandner, W., & Figer, D. F. 2002, *A&A*, 394, 459
- Stothers, R. B. 1992, *ApJ*, 392, 706
- Strader, J., Caldwell, N., & Seth, A. C. 2011, *AJ*, 142, 8
- Tan, J. C., Krumholz, M. R., & McKee, C. F. 2006, *ApJ*, 641, L121
- Testi, L., & Sargent, A. I. 1998, *ApJ*, 508, L91
- Thies, I., & Kroupa, P. 2007, *ApJ*, 671, 767
- Thies, I., & Kroupa, P. 2008, *MNRAS*, 390, 1200
- Thies, I., Kroupa, P., & Theis, C. 2005, *MNRAS*, 364, 961
- Thies, I., Kroupa, P., Goodwin, S. P., Stamatellos, D., & Whitworth, A. P. 2010, *ApJ*, 717, 577
- Tilley, D. A., & Pudritz, R. E. 2004, *MNRAS*, 353, 769
- Tilley, D. A., & Pudritz, R. E. 2007, *MNRAS*, 382, 73
- Tout, C. A., Livio, M., & Bonnell, I. A. 1999, *MNRAS*, 310, 360
- Treyer, M., Wyder, T., Neill, J., Seibert, M., & Lee, J. (eds.), 2011, *ASP Conf. Ser.* 440, UP2010: Have Observations Revealed a Variable Upper End of the Initial Mass Function? (San Francisco, CA: ASP)
- Tsujimoto, T. 2011, *ApJ*, 736, 113
- Turner, J. L. 2009, Extreme star formation, in *Astrophysics in the Next Decade*, eds. H. A. Thronson, M. Stiavelli, & A. Tielens (Dordrecht/London: Springer), 215
- Ulvestad, J. S. 2009, *AJ*, 138, 1529
- Vanbeveren, D. 1982, *A&A*, 115, 65
- Vanbeveren, D. 2011, arXiv:astro-ph/1109.6497
- van Dokkum, P. G. 2008, *ApJ*, 674, 29
- van Dokkum, P. G., & Conroy, C. 2010, *Nature*, 468, 940
- Vazdekis, A., Casuso, E., Peletier, R. F., & Beckman, J. E. 1996, *ApJS*, 106, 307
- Vazdekis, A., Peletier, R. F., Beckman, J. E., & Casuso, E. 1997, *ApJS*, 111, 203
- Veltchev, T. V., Klessen, R. S., & Clark, P. C. 2011, *MNRAS*, 411, 301
- Vesperini, E., & Heggie, D. C. 1997, *MNRAS*, 289, 898
- Vesperini, E., McMillan, S. L. W., & Portegies Zwart, S. 2009, *ApJ*, 698, 615
- Vogt, S. S., Butler, R. P., & Marcy, G. W., et al. 2002, *ApJ*, 568, 352
- von Hippel, T., Gilmore, G., Tanvir, N., Robinson, D., & Jones, D. H. P. 1996, *AJ*, 112, 192
- Wang, P., Li, Z.-Y., Abel, T., & Nakamura, F. 2010, *ApJ*, 709, 27
- Weidemann, V. 1990, in *NATO ASIC Proc. 305, Baryonic Dark Matter*, eds. D. Lynden-Bell & G. Gilmore (Dordrecht/Boston: Kluwer), 87
- Weidemann, V., Jordan, S., Iben, I. J., & Casertano, S. 1992, *AJ*, 104, 1876
- Weidner, C., & Kroupa, P. 2004, *MNRAS*, 348, 187
- Weidner, C., & Kroupa, P. 2005, *ApJ*, 625, 754
- Weidner, C., & Kroupa, P. 2006, *MNRAS*, 365, 1333
- Weidner, C., Kroupa, P., & Larsen, S. S. 2004, *MNRAS*, 350, 1503
- Weidner, C., Kroupa, P., & Maschberger, T. 2009, *MNRAS*, 393, 663
- Weidner, C., Kroupa, P., & Bonnell, I. A. 2010, *MNRAS*, 401, 275
- Weidner, C., Kroupa, P., & Pflamm-Altenburg, J. 2011, *MNRAS*, 412, 979
- Weigelt, G., & Baier, G. 1985, *A&A*, 150, L18
- Weisz, D. R., Johnson, B. D., & Johnson, L. C., et al. 2012, *ApJ*, 744, 44
- White, R. J., & Basri, G. 2003, *ApJ*, 582, 1109
- Whitworth, A. P., & Zinnecker, H. 2004, *A&A*, 427, 299
- Wilkins, S. M., Hopkins, A. M., Trentham, N., & Tojeiro, R. 2008, *MNRAS*, 391, 363
- Winitzki, S. 2003, in *LNCS 2667/2003, Computational Science and Its Applications ICCSA 2003*, eds. V. Kumar et al. (Berlin/Heidelberg: Springer), 780–789
- Wojtak, R., Hansen, S. H., & Hjorth, J. 2011, *Nature*, 477, 567
- Wolfire, M. G., & Cassinelli, J. P. 1987, *ApJ*, 319, 850
- Wuchterl, G., & Klessen, R. S. 2001, *ApJ*, 560, L185
- Wuchterl, G., & Tscharnuter, W. M. 2003, *A&A*, 398, 1081
- Wünsch, R., Silich, S., Palouš, J., Tenorio-Tagle, G., & Muñoz-Tuñón, C. 2011, *ApJ*, 740, 75
- Yasui, C., Kobayashi, N., Tokunaga, A. T., Terada, H., & Saito, M. 2008, *ApJ*, 675, 443
- Zinnecker, H. 1984, *MNRAS*, 210, 43
- Zinnecker, H. 2011, in *ASP Conf. Ser.* 440, eds. M. Treyer, T. Wyder, J. Neill, M. Seibert, & J. Lee, 3
- Zinnecker, H., & Yorke, H. W. 2007, *ARA&A*, 45, 481
- Zucker, S., & Mazeh, T. 2001, *ApJ*, 562, 1038

5 The Galactic Nucleus

Fulvio Melia

Department of Physics, Steward Observatory, and the Applied Math Program, The University of Arizona, Tucson, AZ, USA

1	<i>Introduction</i>	244
2	<i>Radio Morphology of the Galactic Nucleus</i>	247
3	<i>X-ray Morphology of the Central Region</i>	250
4	<i>The Supermassive Black Hole</i>	256
5	<i>The Central Star Cluster</i>	260
6	<i>The Environment Surrounding Sagittarius A*</i>	263
7	<i>Strong Field Physics</i>	266
	<i>References</i>	269

Abstract: Exciting new broadband observations of the galactic nucleus have placed the heart of the Milky Way under intense scrutiny in recent years. This has been due in part to the growing interest from theorists motivated to study the physics of black hole accretion, magnetized gas dynamics, and unusual star formation. The center of our Galaxy is now known to harbor the most compelling supermassive black hole candidate, weighing in at 3–4 million solar masses. Its nearby environment is comprised of a molecular dusty ring, clusters of evolved and young stars, diffuse hot gas, ionized gas streamers, and several supernova remnants. This chapter will focus on the physical makeup of this dynamic region and the feasibility of actually imaging the black hole’s shadow in the coming decade with mm interferometry.

Keywords: Black hole imaging; Black hole physics; Gas dynamics; Interferometry; Star formation; Stellar kinematics; Supernova remnants

1 Introduction

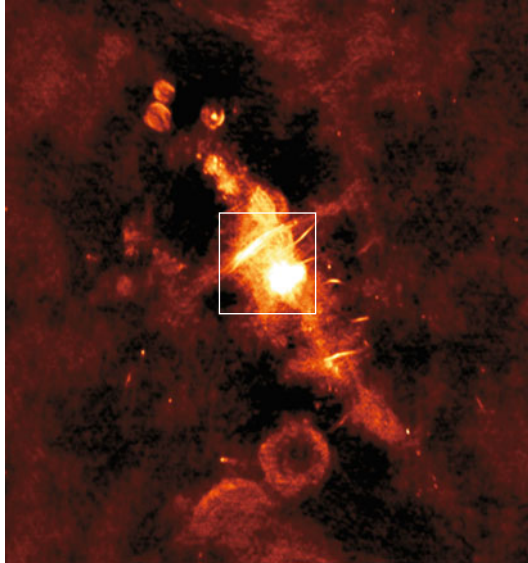
The identification of the Galaxy’s structure and its central region began in the early decades of the nineteenth century, prior to which the Milky Way was thought to encompass the entire universe, with the solar system at its nucleus. The earliest indication that the center of the Milky Way might be far from the solar system was a peculiarity observed in the distribution of globular clusters – gravitationally bound aggregates of thousands to millions of stars, spread over a volume several hundred light-years in diameter. Characterized by high central densities and a round structure, globular clusters apparently formed early in the history of the universe.

John Herschel (1792–1871) noticed in the 1830s that a large number of these clusters occurred in a relatively small portion of the sky, mainly in the direction of Sagittarius. Almost 100 years later, Harlow Shapley (1885–1972) correctly interpreted this unusual distribution while studying what he thought were Cepheid variables sprinkled among the aggregated stars.

A Cepheid is a young star of several solar masses and roughly 10,000 solar luminosities, whose brightness changes repetitively with a period correlated to its intrinsic luminosity. Thus, measuring the repetition cycle also provides an unmistakable determination of its absolute brightness, making the Cepheid a reliable standard candle for assessing the cosmic distance scale.

Assuming a symmetric spatial distribution of globular clusters about the galactic center, Shapley concluded that the globular clusters formed a halo around a flat disk-shaped body with a diameter of 300,000 light-years and a centroid some 50,000 light-years away. It was realized only later that the standard candles he had been observing were not Cepheids at all but rather RR Lyrae variables – stars similar to Cepheids, though fainter. The Milky Way is now thought to be 100,000 light-years across, and the distance to its center is only 20,000–30,000 light-years.

A century later, it became increasingly clear that the location of the galactic center could not be discussed in isolation from the dominant radio source in this region, known as Sagittarius A (see [▶ Figs. 5-1](#) and [▶ 5-2](#)). The strong maximum of radio continuum emission in the constellation Sagittarius was first recognized as a discrete source in 1951 by Jack H. Piddington (1910–1997) and Harry C. Minnett. Because the bright emission originated from the center of the Milky Way, as indicated by optical observations, it was assumed that Sagittarius A should indeed be at the galactic nucleus. For this reason, Sagittarius A was subsequently used to define the zero of longitude in the revised system of galactic coordinates (see Blaauw et al. 1960).



■ Fig. 5-1

At 90 cm, the galactic center is one of the brightest and most intricate regions of the sky. This VLA radio image spans an area of about 1,000 light-years on each side, revealing a rich morphology produced by supernova remnants (the circular features), wispy synchrotron filaments, and highly ionized hydrogen gas. The galactic plane in this image runs from the upper left to the lower right. A schematic diagram of the extended sources seen here is shown in ● Fig. 5-2. The region within the central box is magnified in ● Fig. 5-3 (Produced at the U.S. Naval Research Laboratory by Dr. N. E. Kassim and collaborators from data obtained with the National Radio Astronomy's Very Large Array Telescope, a facility of the National Science Foundation operated under cooperative agreement with Associated Universities, Inc. This image originally appeared in LaRosa et al. 2000)

More recently, remarkable proper motion data acquired over a decade of observations have facilitated the measurement of the Galaxy's central mass distribution down to a field as small as 5 light-days (McGinn et al. 1989; Rieke and Rieke 1989; Sellgren et al. 1990; Haller et al. 1996; Genzel et al. 1996; Eckart and Genzel 1996; Ghez et al. 1998, 2003; Schodel et al. 2002). Two clusters of massive and evolved stellar systems orbit with increasing speed toward the center, where a concentration of dark mass ($\sim 3\text{--}4 \times 10^6 M_{\odot}$) dominates the gravitational potential in a region no bigger than 0.015 pc, or roughly 800 astronomical units. The stars move on Keplerian orbits consistent with a supermassive black hole, called Sagittarius A* (a radio point source within the bigger diffuse emitting region Sagittarius A), as the likely manifestation of this dark matter. Its inferred mass is one of the most accurately known for such an object.

But the galactic nucleus harbors a far more complex tapestry of mutually interacting components than this simple description would imply. The inner 20 light-years or so also contains an enshrouding cluster of evolved stars, an assembly of young stars, and molecular and gas clouds, all accentuated by a powerful supernova remnant, known as Sagittarius A East.¹ Not

¹The heart of the Milky Way lies in the direction of the constellation Sagittarius, close to the border with the neighboring constellation Scorpius. Celestial objects and features tend to be named after the constellation in which they are found, so the galactic center is said to lie in the Sagittarius A complex, and gaseous structure within it is called, for example, Sagittarius A East and Sagittarius A West.

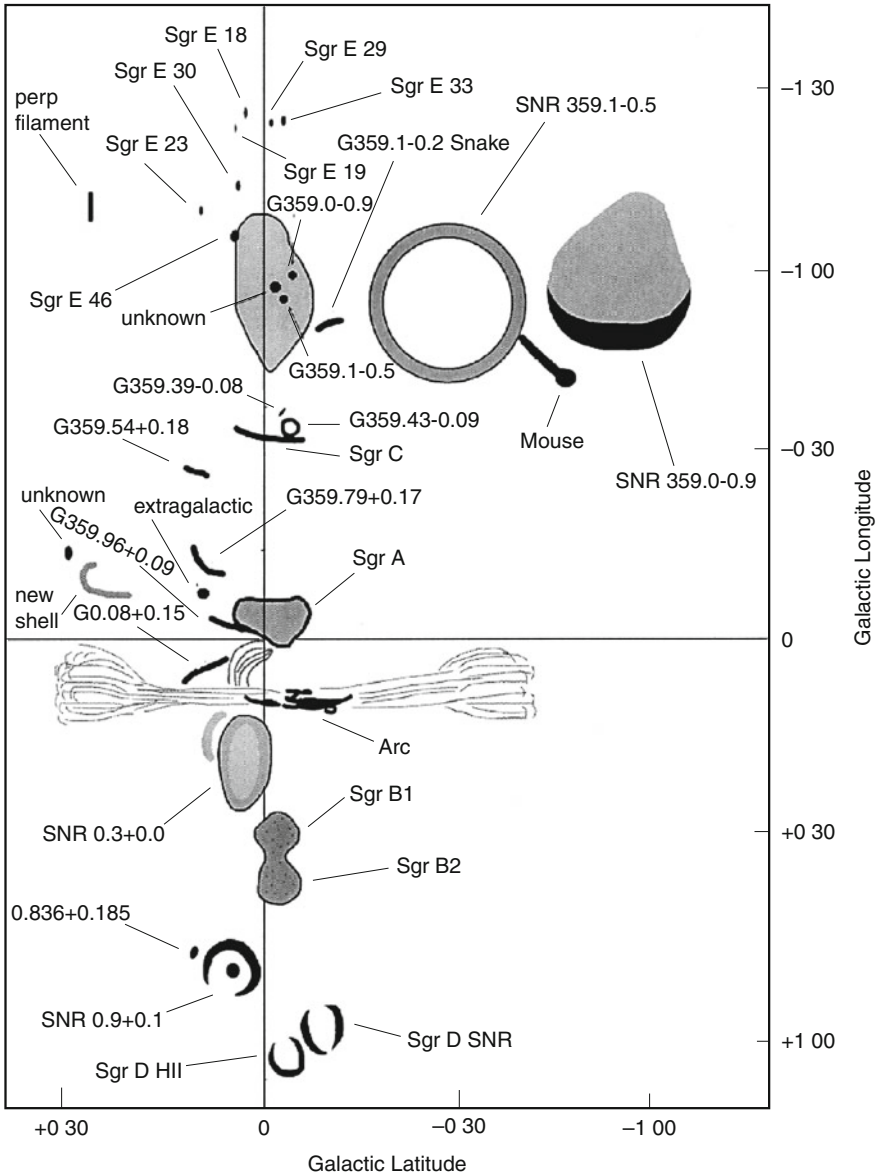


Fig. 5-2
 Schematic diagram of the extended sources shown in the 90-cm image of the galactic center (Fig. 5-1). Here, the galactic plane is vertical (From LaRosa et al. 2000)

surprisingly, many link the center of our Galaxy to the more extensive class of active galactic nuclei (AGN), in which a supermassive black hole functions as the principal agent behind much of the observed dynamical and radiative behavior of the host galaxy's core. For this reason, developing a consistent picture of the primary interactions between the various constituents at the galactic center improves the picture of AGN activity in a broader context.

This chapter examines the principal components residing within the Galaxy's inner core and their overall morphology, revealed primarily through the power of modern X-ray and radio telescopes. Today, very little doubt remains concerning the nature of Sagittarius A*, particularly if its radio characteristics are indeed closely associated with the dark matter concentrated within only a fraction of a parsec of the center.

2 Radio Morphology of the Galactic Nucleus

☛ *Figure 5-1* shows a wide-field, high-resolution 90-cm image centered on Sagittarius A, covering an area of $4^\circ \times 5^\circ$ with an angular resolution of $43''$. (At the ~ 8 kiloparsec distance to the galactic center, $1''$ is approximately 0.04 pc.) This map, produced by LaRosa et al. (2000), is based on archival data originally acquired and presented by Pedlar et al. (1989) and Anantharamaiah et al. (1991), who observed the galactic center with the VLA 333 MHz system in all four array configurations between 1986 and 1989. A schematic diagram in galactic coordinates of the extended sources seen in the 90-cm image is shown in ☛ *Fig. 5-2*.

With the exception of the Sagittarius A complex centered on Sagittarius A*, nearly all of the sources in ☛ *Fig. 5-1* are detected in emission, providing a view of the large-scale radio structure in the galactic nucleus. Note, however, that of the 78 small-diameter ($< 1'$) sources concentrated toward the galactic plane, about half have steep spectra ($\alpha \approx -0.8$) and are therefore probably extragalactic. Within the central $15'$ (or roughly 37 pc), the most notable structure is the Sagittarius A complex, consisting of the compact nonthermal source Sagittarius A*, surrounded by an orbiting spiral of thermal gas known as Sagittarius A West.² Along the same line of sight lies the nonthermal shell source known as Sagittarius A East, which appears to be the remnant of an energetic explosion. Sagittarius A West is seen in absorption against the background of Sagittarius A East, indicating that the latter must lie behind the former.

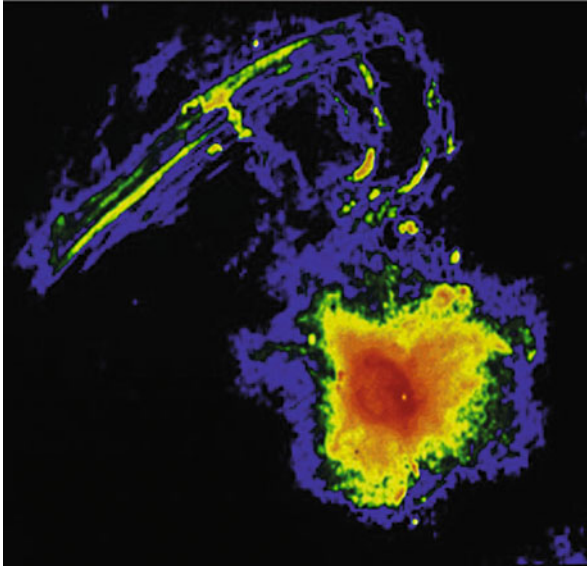
Some $15' - 20'$ (or ~ 50 pc in projection) north of Sagittarius A is located the galactic-center arc. First resolved into a large number of narrow filaments by Yusef-Zadeh et al. (1984), they show strong polarization with no line emission and are therefore nonthermal synchrotron sources, probably magnetic flux tubes flushed with relativistic electrons. Several other (isolated) filaments within the central half degree also contribute to the nonthermal magnetic structure, and most are oriented perpendicular to the galactic plane.

Stars are not visible in this image, but their births and deaths impact the entire region, as evidenced by the presence of supernova remnants, such as Sgr D and SNR 0.9 + 0.1. Giant molecular clouds (such as Sgr B1 and Sgr B2) are regions of star formation and become discernible when newborn stars heat the surrounding gas and make it shine in the radio. All in all, this radio continuum view, together with observations at mm, infrared, and X-ray wavelengths (see next sub-section), points to the galactic center as constituting a weak, Seyfert-like nucleus that sometimes also displays mild outbursts of active star formation.

The region bounded by the box in ☛ *Fig. 5-1* is shown at 20 cm in ☛ *Fig. 5-3*, a radio continuum image spanning the inner 50-pc-by-50-pc portion of the Galaxy. On this level, the distribution of hot gas within the Sagittarius A complex displays an even richer morphology than at 90 cm, with the evident coexistence of both thermal and nonthermal components.³

²A description of this structure may be found in Ekers et al. (1983), and Lo and Claussen (1983).

³See also Yusef-Zadeh and Morris (1987) and Pedlar et al. (1989).



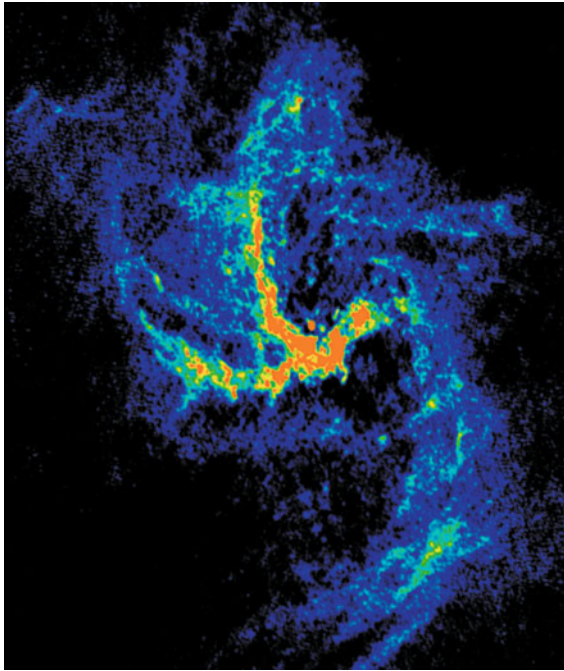
■ Fig. 5-3

Magnified view of the central, bright region in [Fig. 5-1](#), spanning a few hundred light-years in either direction (The galactic plane runs from the *upper left* to the *bottom right*, as in [Fig. 5-1](#)). This VLA radio image shows the intensity of radiation at 20 cm, produced mostly by magnetized, hot gas between the stars. The system of narrow filaments (some wrapped around each other) have a width of about 3 light-years, and tend to be oriented perpendicular to the galactic plane. The bright supernova remnant surrounding the center of our Galaxy in the *lower right* quadrant is magnified further in [Fig. 5-4](#). The spot in the middle of the red spiral identifies the radio source known as Sagittarius A*, believed to be the radiative manifestation of the supermassive black hole at the center of our Galaxy (Image courtesy of F. Yusef-Zadeh, and the National Radio Astronomy Observatory/Associated Universities, Inc.)

Sagittarius A East is the diffuse ovoid region to the lower right in [Fig. 5-3](#), surrounding (in projection) a spiral-like pattern in red, which is Sagittarius A West. The central spot in this structure identifies Sagittarius A*, which is coincident with the concentration of dark matter inside 0.015 pc.

At a wavelength of 6 cm (see [Fig. 5-4](#)), Sagittarius A West appears as a three-armed spiral consisting of highly ionized gas radiating a thermal continuum. Each arm in the spiral is about 3 light-years long, but this structure may be merely a superposition of gas streamers seen in projection. At a distance of 3 light-years from the center, the plasma moves at a velocity of about 105 km s^{-1} , requiring a mass concentration of just over $3.5 \times 10^6 M_{\odot}$ inside this radius. The hub of the gas spiral corresponds to the very bright radio source Sagittarius A*, the dynamical center of our Galaxy.

The central 2 light-year \times 2 light-year portion of Sagittarius A West is shown at 2 cm in [Fig. 5-5](#). Sagittarius A West probably derives its heat from the central distribution of bright stars, rather than from a single point source, such as Sagittarius A*. Some hot, luminous stars



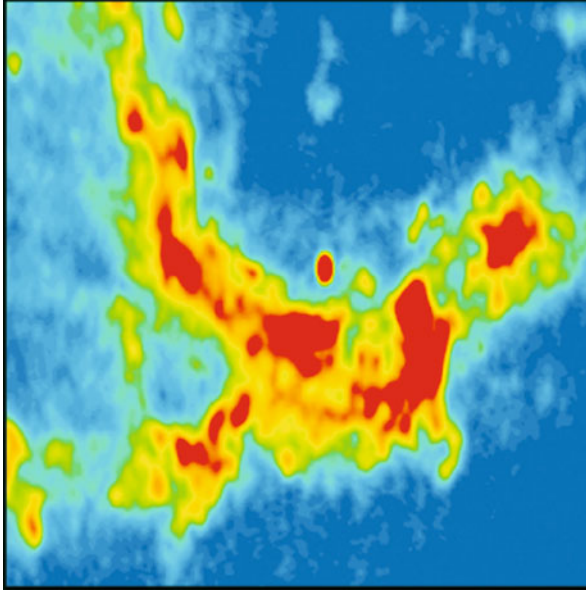
■ Fig. 5-4

Magnified view at 6 cm of the (red) spiral structure seen in [Fig. 5-3](#). Each of the “arms” is about 3 light-years in length; these may constitute either a real spiral pattern or perhaps a superposition of independent gas flows into the center. The latest measurements indicate that this gas is moving about the nucleus with a velocity as high as $1,000 \text{ km s}^{-1}$. The central region is magnified further in [Fig. 5-5](#) (Image courtesy of F. Yusef-Zadeh at Northwestern University, and the National Radio Astronomy Observatory)

are thought to have been formed as recently as a few million years ago, though it is not known yet whether these particular stars formed within the gas streamer or just happen to lie along the line of sight.

On a scale of $\sim 3 \text{ pc}$, Sagittarius A West orbits about the center within a large central cavity, surrounded by a gaseous and dusty circumnuclear ring (see Becklin et al. 1982, and Davidson et al. 1992). [Figure 5-6](#) shows a radio-wavelength image of ionized gas at 1.2 cm (due to free-free emission) superimposed on the distribution of hydrogen cyanide (HCN), which traces the molecular gas. This clumpy molecular ring has an inferred mass of more than $10^4 M_{\odot}$, and rotates around a concentrated cluster of hot stars, known as IRS 16, with a velocity of about 110 km s^{-1} , according to Gusten et al. (1987) and Jackson et al. (1993).

The IRS 16 complex consists of about two dozen blue stellar components at $2 \mu\text{m}$ and appears to be the source of a strong wind with velocity on the order of 700 km s^{-1} and an inferred mass loss rate of $4 \times 10^{-3} M_{\odot} \text{ year}^{-1}$. Most of the far infrared luminosity of the circumnuclear ring is due to this cluster of hot, helium emission-line stars, which bathe the central cavity with



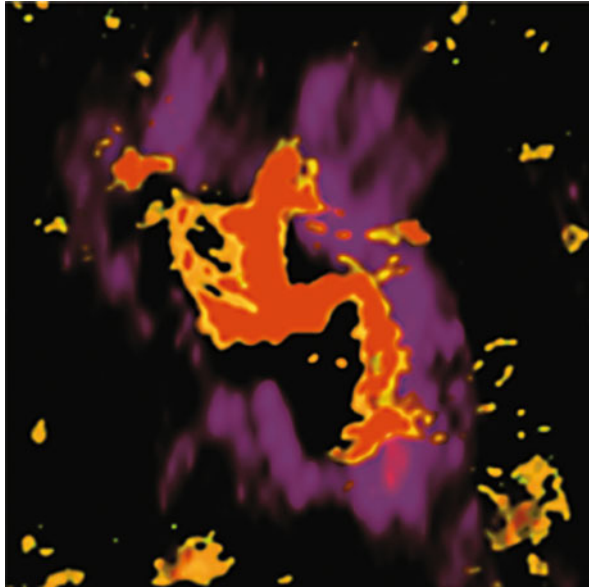
■ Fig. 5-5

At 2 cm, the innermost 2 light-year \times 2 light-year region of **Fig. 5-4** is dominated by the central portion of the spiral pattern of Sagittarius A West and a bright point-like source known as Sagittarius A*, near the middle of the image. To the north of Sagittarius A*, the cometary-like feature (in *light blue* against the *dark blue* background) is associated with the luminous red giant star IRS 7. The gas blown upward from its envelope provides evidence of a strong wind emanating from the region near the supermassive black hole. The distance between Sagittarius A* and the red giant is $\sim 3/4$ light-year (Image courtesy of F. Yusef-Zadeh at Northwestern University, and the National Radio Astronomy Observatory)

ultraviolet radiation, heating the dust and gas up to 8 pc from the center of the Galaxy. These blue stars are themselves embedded within a cluster of evolved and cool stars with a radial density distribution r^{-2} from the dynamical center. However, unlike the distribution of evolved cluster members, which extend over the central 500 pc of the galactic bulge, the hot stars in IRS 16 are concentrated only within the inner parsecs (Hall et al. 1982; Geballe et al. 1987; Allen et al. 1990).

3 X-ray Morphology of the Central Region

An equally interesting view of the galactic nucleus emerges with progressively sharper images of this region in the X-ray band. X-ray emission has been observed on all scales, from structure extending over kiloparsecs down to a fraction of a light-year, with contributions from thermal and nonthermal, point-like, and diffuse sources. **Figure 5-7**, which shows the 1.5 keV map produced with ROSAT (Snowden et al. 1997), contains evidence for a large-scale outflow of hot gas from the nucleus. The hollow-cone-shaped soft X-ray feature on either side of the galactic



■ Fig. 5-6

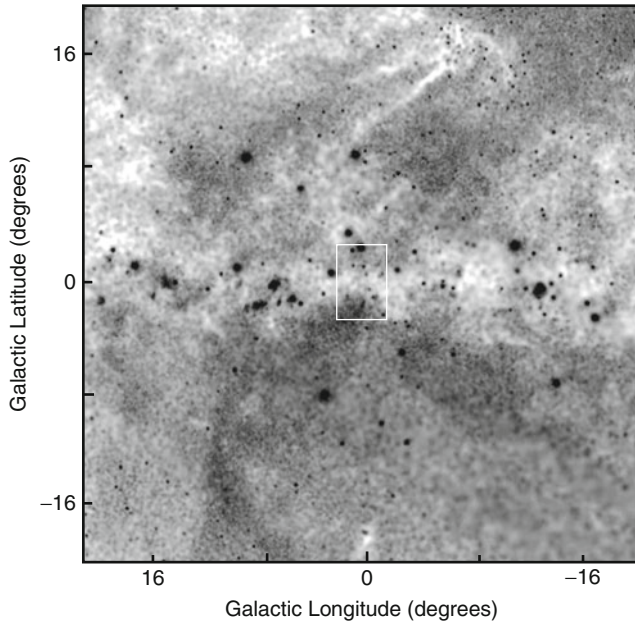
A radio image of ionized gas (Sgr A West) at $\lambda = 1.2$ cm, with its three-arm structure (*orange*) superimposed on the distribution of HCN emission (*violet*), providing evidence for the presence of a torus of dusty gas in orbit about the central source of gravity, Sagittarius A*. The dust in this ring shines by converting ultraviolet light into an infrared glow. At the distance to the galactic center, this image corresponds to a size of approximately 4 pc on each side (Image courtesy of F. Yusef-Zadeh at Northwestern University, M. Wright at the Radio Astronomy Laboratory, University of California at Berkeley, and the National Radio Astronomy Observatory)

plane resembles the morphology seen in nearby galaxies with active nuclear star formation. The efflux of plasma responsible for this structure accounts for much of the diffuse soft X-ray background in the Milky Way. The presence of various spectral features, particularly the 6.7 keV Fe XXV $K\alpha$ line detected with ASCA, suggests further that a large fraction of this gas is so hot⁴ that confinement due to gravity is not feasible, though observations with *Chandra* have more recently suggested a refinement to this global conclusion (more on this below).

The magnified view of the central $3^\circ \times 3^\circ$ shown in [► Fig. 5-8](#) reveals additional evidence for the expulsion of hot matter from the nucleus, in the form of a prominent, bright soft X-ray plume that apparently connects the galactic center to the large-scale X-ray structure hundreds of parsecs above and below the galactic plane.

But the most detailed X-ray view of the galactic center has been provided by *Chandra*'s Advanced CCD Imaging Spectrometer (ACIS) detector, which combines the wide-band sensitivity and moderate spectral resolution of ASCA and Beppo-SAX with the much higher spatial resolution ($\sim 0.5''$ – $1''$) of *Chandra*'s High-Resolution Mirror Assembly (HRMA). The central rectangular box oriented along the galactic plane in [► Fig. 5-8](#) outlines the field mapped out in the 1–8 keV range by the most complete *Chandra* survey to date. This study consists of 30

⁴The ASCA Fe-line observations apparently require a temperature as high as $\sim 10^8$ K. See Koyama et al. (1996).



■ Fig. 5-7

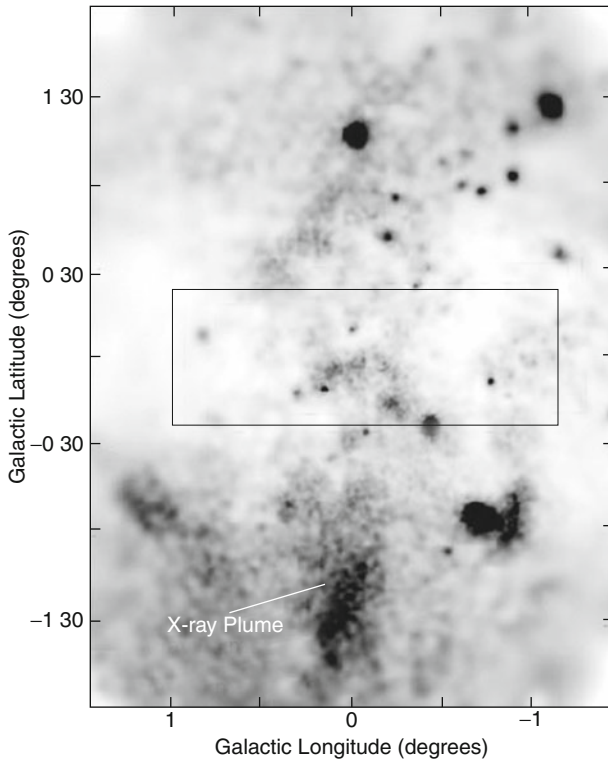
ROSAT all-sky survey of the inner $40^\circ \times 40^\circ$ region of the Galaxy in the ~ 1.5 keV band ($1^\circ \approx 144$ pc). The central box marks the region targeted by ROSAT pointed observations and is shown magnified in [Fig. 5-8](#) (Image courtesy of S. L. Snowden at MPE, the Goddard Space Flight Center, and NASA)

separate pointings, all taken in July 2001; a mosaic of these observations is shown in [Fig. 5-9](#), covering a field of view $\sim 2^\circ \times 0.8^\circ$ centered on Sagittarius A. The saw-shaped boundaries of this map, plotted in galactic coordinates, result from a specific roll angle of the observations (see Wang et al. 2002).

Chandra's high spatial resolution allows for a separation of the discrete sources from the diffuse X-ray components pervading the galactic-center region. These observations have led to a detection of roughly 1,000 discrete objects within the inner $2^\circ \times 0.8^\circ$, whose number and spectra indicate the presence of numerous accreting white dwarfs, neutron stars, and solar-size black holes. As many as half of these discrete objects could be luminous background active galactic nuclei. Most of the other sources have a luminosity $\sim 10^{32}$ – 10^{35} ergs s^{-1} in the 2–10 keV band.

A fundamental question that motivated the *Chandra* survey concerns the relative contribution of the point-source and diffuse components to the overall X-ray emission from the center of the Milky Way. For example, earlier observations with ASCA (Tanaka et al. 2000) had implied that the ubiquitous and strong presence of the He-like Fe $K\alpha$ line (at ~ 6.7 keV) throughout the central region required the existence of large quantities of $\sim 10^8$ K gas – a situation that is very difficult to explain on physical grounds.

A direct comparison between the accumulated point-source spectrum within the central region with that of the diffuse emission (see [Fig. 5-10](#)) reveals a distinct emission feature centered at ~ 6.7 keV (with a Gaussian width of ~ 0.09 keV) in the former but not the latter. The characteristics of this feature agree with those inferred previously with ASCA.

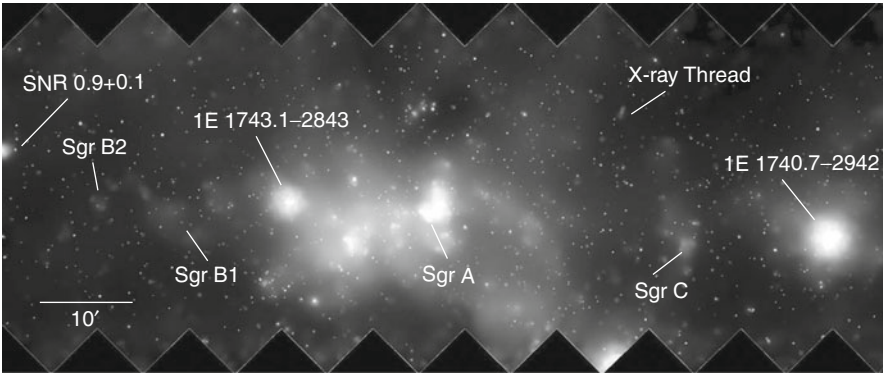


■ Fig. 5-8

Close-up mosaic of the central $3^\circ \times 3^\circ$ of the Galaxy, constructed with ROSAT PSPC observations in the highest energy band (0.5–2.4 keV). The bright, soft X-ray plume apparently connects the central region to the southern large-scale X-ray cone some 300 pc away from the plane. The plume is the most prominent and coherent vertical diffuse soft X-ray feature seen at the galactic center; it may represent the hot gas outflow from the nucleus into the surrounding halo. The central rectangular box oriented parallel to the galactic plane outlines the field mapped out by the more recent Chandra survey, shown in ▶ Fig. 5-9 (Image courtesy of L. Sidoli and S. Mereghetti at INAF-IASF Milano, and T. Belloni at INAF-Osservatorio di Brera)

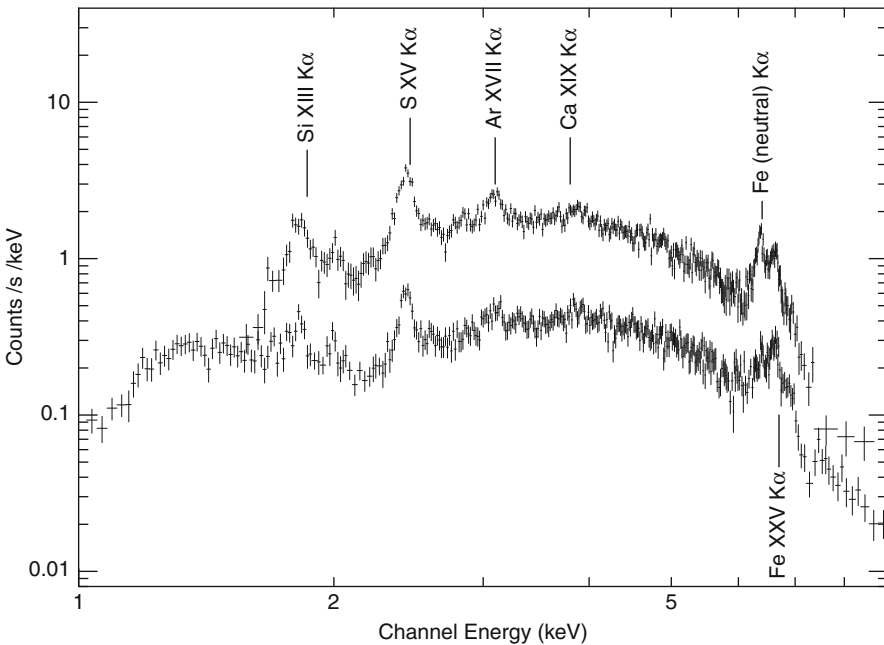
The high-resolution *Chandra* measurements seem to have resolved the issue of how the He-like Fe $K\alpha$ line is produced – this emission is typical of X-ray binaries containing white dwarfs, neutron stars, or black holes, particularly during their quiescent state. Rather than being attributed to the diffuse emission, the He-like Fe $K\alpha$ line is instead found largely due to these discrete X-ray source populations.

Note, however, that the line emission from ions such as S XV, Ar XVII, and Ca XIX is quite prominent in the diffuse X-ray spectrum, which together with the weaker He-like Fe-line points instead to the presence of an optically thin thermal plasma with a characteristic temperature of $\sim 10^7$ K – typical of young supernova remnants.



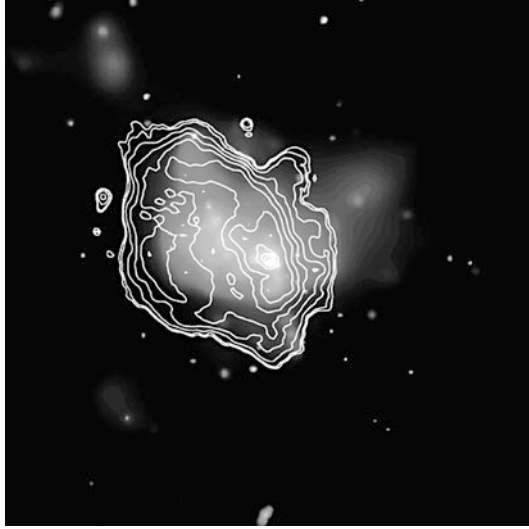
■ Fig. 5-9

Mosaic covering an $\sim 2^\circ \times 0.8^\circ$ band in galactic coordinates centered at $(l'', b'') = (-0.1^\circ, 0^\circ)$. Based on 30 separate observations made in July 2001, this image shows intensity in three different energy bands: 1–3 keV (*brightest*), then 3–5 keV, and finally 5–8 keV (*faintest*) (Image courtesy of Q. Daniel Wang at the University of Massachusetts, Amherst, and NASA)



■ Fig. 5-10

The Chandra spectrum of the diffuse X-ray flux enhancement above the surrounding background (*upper curve*) is shown in comparison with that of the accumulated point-source radiation (*lower curve*). This comparison seems to settle the issue of how the He-like Fe-line is produced (see text) (From Wang et al. 2002)



■ Fig. 5-11

The smoothed broadband X-ray intensity map (1.5–7.0 keV) from Chandra overlaid with radio contours from a 20 cm VLA image of Sagittarius A (see ● Fig. 5-3). The outer oval-shaped radio structure is associated with synchrotron emission from the shell-like nonthermal radio source Sagittarius A East (Image from Maeda et al. 2002)

Even so, it is difficult to avoid the fact that the overall spectrum of diffuse X-ray emission (● Fig. 5-10) at the galactic nucleus is considerably harder than expected for a thermal component alone. Nearly half of the detected diffuse emission in the 5–8 keV band is due to the Fe 6.4 keV line, part of which is likely due to the fluorescent radiation from discrete sources. Images such as ● Fig. 5-9 indicate that the distribution of the line emission tends to be correlated with lumpy dense molecular material. The problem is that the known population of bright X-ray objects in the galactic center region is not sufficient to produce this fluorescence. Instead, it is likely that certain X-ray sources – possibly even the supermassive black hole itself – may have varied greatly in the past, so that their averaged luminosity was several orders of magnitude higher than today. Much of the present 5–8 keV diffuse emission could then be due to this past discrete-source irradiation of the molecular clouds, producing the scattered/fluoresced photon field observed now (see Murakami et al. 2001; Fromerth et al. 2001).


Within the 20-pc-by-20-pc region surrounding Sagittarius A* (see ● Fig. 5-11), the primary origin of the high-energy shroud appears to be Sagittarius A East, the nonthermal radio source with a supernova-like morphology located near the galactic center (see also ● Fig. 5-3). Sagittarius A East is elongated along the galactic plane with a major axis of length 10.5 pc and a center displaced from the apparent dynamical nucleus by 2.5 pc in projection toward negative galactic latitudes.

Sagittarius A East has been detected at 1,720 MHz, the transition frequency of OH maser emission. In general, the detection of this line establishes the presence of shocks at the interface between the supersonic outflow and dense molecular material with which the remnant is interacting. Several maser spots with velocities $\approx 50 \text{ km s}^{-1}$ have been resolved in the region


where this remnant is interacting with the dense molecular cloud known as M–0.02–0.07, at the southeastern boundary. The detection of these OH masers is a principal reason behind the identification of Sagittarius A East as the remnant of a powerful explosion.

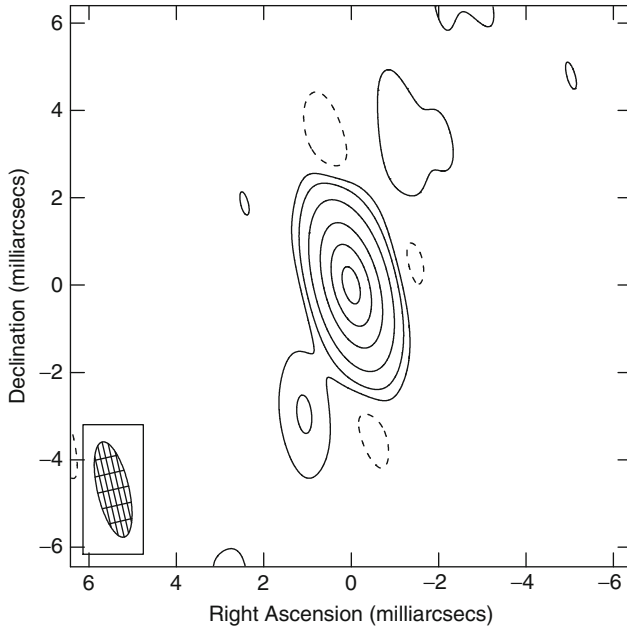
The X-ray spectrum of Sagittarius A East contains strong $K\alpha$ lines from highly ionized ions of S, Ar, Ca, and Fe, for which a simple isothermal model yields an electron temperature ~ 2 keV. The inferred metallicity is overabundant by a factor of four compared with solar values, concentrated toward the middle. Maeda et al. (2002) conclude from this that Sagittarius A East is probably the result of a Type II supernova explosion, with a 13–20 M_{\odot} main-sequence progenitor, and that the combination of its radio and X-ray properties classifies it as a metal-rich “mixed morphology” remnant. However, the size of the Sagittarius A East shell is the smallest known for this category of sources, implying that the ejecta have been expanding into a uniquely dense interstellar medium. For a 10,000-year-old structure, the implied ambient density is $\sim 10^3$ cm^{-3} , fully consistent with the observed properties of the 50 km s^{-1} M–0.02–0.07 molecular cloud, into which Sagittarius A East is apparently expanding.

4 The Supermassive Black Hole

The central point-like object, coincident with Sagittarius A* in  Fig. 5-5, was discovered on February 13 and 15, 1974, by Balick and Brown (1974), who reported this “detection of strong radio emission in the direction of the inner 1-pc core of the galactic nucleus” later that year. Eventually, it would be viewed as the most unusual source in the Galaxy. The novelty that permitted them to distinguish point-like objects from the overall radio emission in the inner 20'' was the newly commissioned 35-km baseline interferometer of the National Radio Astronomy Observatory, consisting of three 26-m telescopes separable by up to 2.7 km and a new 14-m telescope located on a mountaintop about 35 km southwest of the other dishes.

The motivation for establishing that the galactic center is active in ways similar to more powerful galactic nuclei had been discussed and developed over the previous three or four years. Sanders and Prendergast (1974) had hypothesized earlier that year that, although now quiescent, the galactic center may once have housed energetic processes like those seen in BL Lac. And in 1971, Lynden-Bell and Rees pointed out that the galactic center should contain a supermassive black hole, perhaps detectable with radio interferometry (Lynden-Bell and Rees 1971). The argument made by Lynden-Bell and Rees was based on the implausibility of starlight alone ionizing the extended thermal source surrounding the central region, not to mention the difficulty of producing a “nuclear wind” with both ionized and neutral material moving at speeds exceeding 200 km s^{-1} . They proposed instead an ultraviolet nonstellar continuum produced by the hypothesized black hole, which presumably also created the observed efflux of mass. Lynden-Bell and Rees’s proposal functioned as an influential catalyst in the early attempts to characterize the new radio source as a black hole phenomenon.

Sagittarius A* is unique among galactic nuclei in that its proximity to Earth allows radio observers to resolve it with very long baseline interferometry (VLBI), though its size is somewhat difficult to determine precisely. The source itself is very compact. With VLBI, the radio telescopes are separated by several thousand kilometers, and unfortunately the major high-frequency VLBI receivers are in the Northern hemisphere, making Sagittarius A* a low-elevation source. This renders the measurements difficult to calibrate because the radio waves must pass through a long column of atmospheric gas.  Figure 5-12 shows a sample image



■ Fig. 5-12

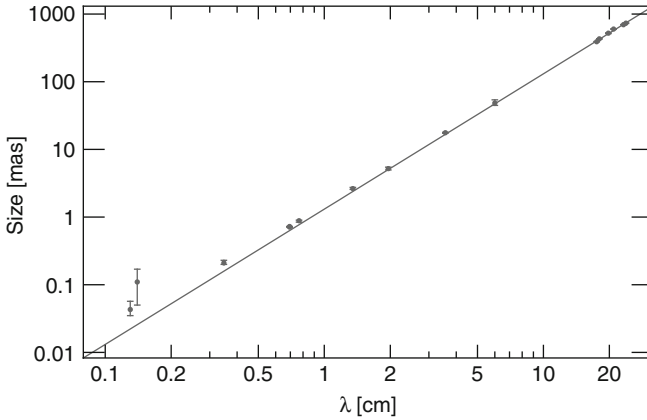
Uniformly weighted image of Sagittarius A* produced with the VLBA at 7 mm. The beam appears in the lower left-hand corner of the diagram. The contours shown here are -0.01 , 0.01 , 0.03 , 0.10 , 0.30 , 0.60 , and 0.90 times the peak intensity of 0.87 Jy/beam. North is up and east is left in this image. The galactic plane runs from the upper left-hand corner to the lower right (Image from Bower and Backer 1998)

obtained at 7 mm with the very large baseline array, illustrating the fact that Sagittarius A* is resolved at mm wavelengths.

Its apparent angular diameter ($\propto \lambda^{+2.0}$ over a large portion of the sampled wavelength range) depends strongly on wavelength, consistent with angular broadening by the scattering of its radio waves in the intervening plasma (see, e.g., Lo et al. 1981, 1985, Backer 1988, and, more recently, Bower and Backer 1998 and Lo et al. 1998). This broadening is very similar to that of OH masers within 0.5° of the galactic nucleus, implying that the diffuse thermal plasma within the central 140 pc (in longitude) is sufficiently turbulent to produce the observed scattering.

The radio size of Sagittarius A* is shown in ● Fig. 5-13 as a function of wavelength. The solid line in this figure represents the scattering law, $\theta_{\text{scat}} = (1.31 \pm 0.02) \text{ mas } (\lambda/\text{cm})^2$, from Bower et al. (2006). The overall size of Sagittarius A* follows this law closely at long wavelengths, but deviates from it in the mm range. It is now understood that, when the size follows this law, the scattering is dominated by fluctuations in the electron density of the interstellar medium. The power spectrum of these fluctuations goes as $k^{-\beta}$, where k is the wavenumber of the irregularities (see Romani et al. 1986). The scattering angle scales as $\lambda^{1+2/(\beta-2)}$, and $[1 + 2/(\beta - 2)] = 2$ when $\beta = 4$.

The intrinsic size is determined by the deconvolution of the observed size θ_{obs} with the extrapolated scattering size θ_{scat} : $\theta_{\text{int}} = (\theta_{\text{obs}}^2 - \theta_{\text{scat}}^2)^{1/2}$. A reasonable fit to the values one gets



■ Fig. 5-13

Measured radio source size (major axis) of Sagittarius A* as a function of the observing wavelength in centimeters. The straight line represents the λ^2 law for broadening of the source structure by scattering along the line-of-sight. At the distance to the galactic center, 1 mas is approximately 8 A.U. Correspondingly, the Schwarzschild radius for a black hole with mass $3.4 \times 10^6 M_\odot$ is about 1/15 A.U. (Image from Falcke et al. 2009)

using this prescription may be written $\theta_{\text{int}} \approx (0.52 \pm 0.03) \text{ mas} \times (\lambda/\text{cm})^{1.3 \pm 0.1}$ (Bower et al. 2006).

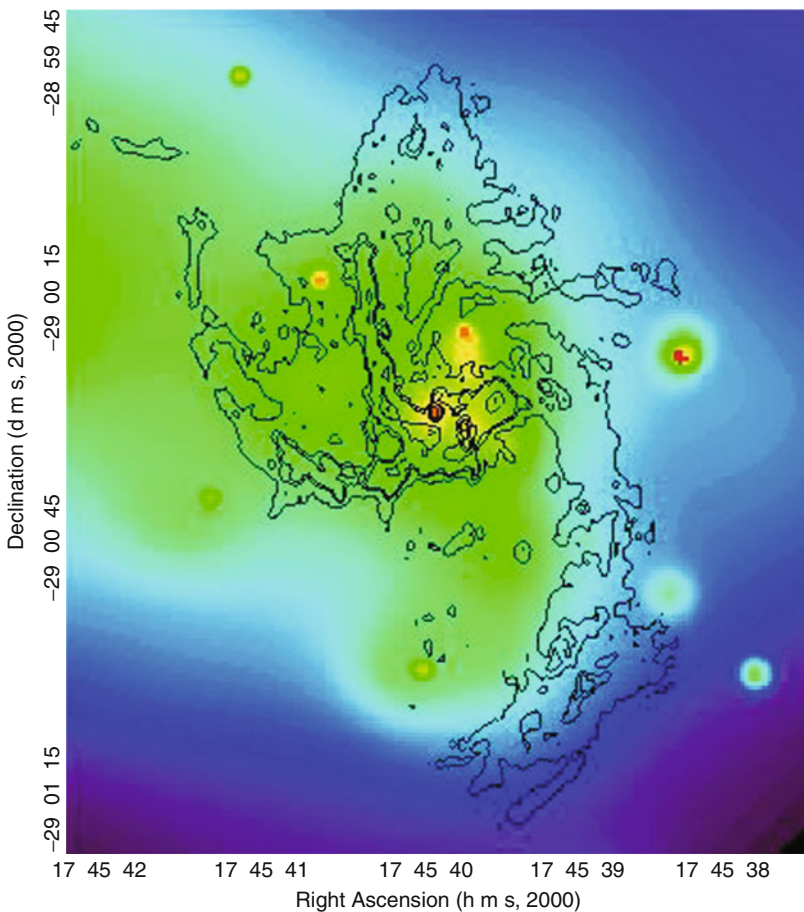
A different technique must be used for measuring Sagittarius A*'s size at mm to sub-mm wavelengths. To date, these observations have been based on refractive scintillation, which sets a *lower* limit to the size of the emission region surrounding the black hole. Unlike the multidish interferometric techniques employed at longer wavelengths, refractive scintillation is associated with observations carried out with single dish telescopes, under the assumption that inhomogeneities in the intervening medium are well understood. To see how this works, imagine staring at a lightbulb, while passing an opaque object across your line of sight. The variation in the intensity of the light would decrease as the size of the eclipsing material is reduced. The clumps in the interstellar medium are about 10^{12} cm wide, so the absence of refractive scintillation in Sagittarius A* at 1.3 and 0.8 mm implies that the intrinsic source must be at least this large (i.e., roughly 0.1 AU) at these wavelengths (Gwinn et al. 1991).

These results together suggest that the emission region surrounding Sagittarius A* is stratified – photons with higher energy are produced at progressively smaller distances from the event horizon. For example, a comparison of the inferred source size at 0.8 mm with that at 7 mm shows that the radiating plasma extends over ≈ 39 Schwarzschild radii in the latter (based on the calculated value of θ_{in}), but only ≈ 1.5 Schwarzschild radii in the former.

Almost certainly, Sagittarius A*'s spectrum extending from radio wavelengths into the infrared is due to synchrotron emission. This conclusion has been strengthened by the recent observation of daily flares from this source, which are associated with enhanced emission lasting anywhere from tens of minutes to several hours. Multi-wavelength observations show that Sagittarius A*'s infrared spectrum is a power law ($F_\nu \propto \nu^\alpha$) with a constant spectral index $\alpha = -0.6 \pm 0.2$. It is notable that α appears to be independent of the intensity, even though the latter can change anywhere from 2 to 30 mJy (Hornstein et al. 2007). In order to change




its infrared intensity without altering α , Sagittarius A* must therefore vary the number of accelerated electrons while leaving their overall energy distribution unchanged.


The measured value of α indicates that the synchrotron emission is optically thin, arising from an electron distribution with number density $N(\gamma_e) \propto \gamma_e^{-p}$, where $p = 1 - 2\alpha \approx 2.2$ and γ_e is the electron Lorentz factor. Further, assuming simple equipartition, it is straightforward to see that the underlying magnetic field must have a strength $B \sim 20$ G, which means that the infrared radiation is produced by ~ 1 GeV electrons (see, e.g., Liu and Melia 2001). How the black hole accelerates these particles is still not understood, though future multi-wavelength observations promise to greatly improve our understanding of this intriguing object.



■ Fig. 5-14

Smoothed X-ray map (0.5–7 keV) of the inner $1.'3 \times 1.'5$ of the Galaxy, overlaid with VLA 6 cm contours corresponding to the intensity map shown in [Fig. 5-4](#). The X-ray emission from Sagittarius A* itself appears as a red dot at $17^{\text{h}}45^{\text{m}}40.0^{\text{s}}, -29^{\circ}00'28''$. Bright diffuse emission from hot gas is visible throughout the region and appears to be produced primarily via wind-wind collisions (From Baganoff et al. 2003)

Additional clues about what may be happening close to Sagittarius A*’s event horizon are now being provided by *Chandra* observations (see, e.g., Baganoff et al. 2003). The (smoothed, false-color) X-ray map of the inner $1.'3 \times 1.'5$ of the Galaxy is shown in  Fig. 5-14. Sagittarius A* appears as a red dot at $17^{\text{h}}45^{\text{m}}40.0^{\text{s}}, -29^{\circ}00'28''$, overlaid with VLA 6 cm contours of the radio intensity in  Fig. 5-4. The western boundary of the brightest diffuse X-ray emission (shown in green) coincides very well with the shape of the Western Arc of thermal emission from Sagittarius A West, whereas the emission on the eastern side continues smoothly into the heart of Sagittarius A East. The Western Arc is thought to be the ionized inner edge of the circumnuclear ring of molecular material orbiting about the galactic center (see  Fig. 5-6), so the morphological similarities between the X-ray and radio features strongly suggest that the brightest X-ray-emitting plasma is confined by the western portion of this ring.


It is important to note that a spatial analysis of the morphology of the source coincident with Sagittarius A* indicates that this source may be slightly extended. *Chandra*’s on-axis high-resolution mirror assembly has a FWHM point-spread function $\approx 0.''5$, whereas the apparent intrinsic size of the central source in  Fig. 5-14 is $\approx 1''$ or about 0.04 pc at the distance to the galactic center. Structure in Sagittarius A* on this scale may be consistent with the radius ($1''\text{--}2''$) at which matter is captured by the black hole and begins its hydrodynamic infall toward the center (see Melia 1994 and Quataert 2002).

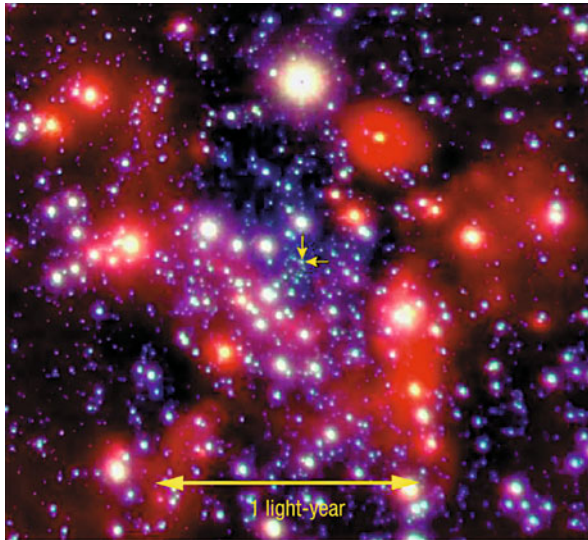
Sagittarius A*’s absorption-corrected luminosity in the 2–10 keV band is measured at $2.4_{-0.6}^{+3.0} \times 10^{33}$ ergs s^{-1} . Sagittarius A* has been difficult to study at high energy because its X-ray luminosity is significantly fainter than that expected of a solar-mass neutron star, let alone a four million solar-mass behemoth. The maximum rate at which a compact object may accrete is dictated by the efficiency with which dissipated gravitational energy is converted into radiation. The interaction between these escaping photons and the infalling plasma retards the flow and can – given the right circumstances – actually lead to a temporary expulsion of the matter rather than its accretion.

The limit on luminosity, known as the Eddington limit, is independent of radius, and is given by the simple expression $L_{\text{Ed}} \equiv 4\pi c GM (m_p + m_e)/\sigma_{\text{T}}$, where M is the black hole mass, m_p and m_e are the proton and electron masses, respectively, and σ_{T} is the Thomson cross section. For a neutron star (with $M \approx 1 M_{\odot}$), $L_{\text{Ed}} \approx 1.3 \times 10^{38}$ ergs s^{-1} , already far greater than Sagittarius A*’s X-ray luminosity estimated by *Chandra*. Indeed, the black hole at the galactic center appears to be radiating at a rate $\sim 10^{10}$ times below the Eddington value for an object with mass $M \approx 3.4 \times 10^6 M_{\odot}$.

This surprising observation represents a serious theoretical problem because estimates of the accretion rate onto Sagittarius A* based on the observed characteristics of its nearby environment would suggest a value much greater than this. Sagittarius A*’s neighborhood will be the focus of the next section; its very low accretion rate will be considered later in this chapter.

5 The Central Star Cluster

Near-infrared, high-spatial resolution imaging observations and spectroscopy of the central stars have provided new insights into their dynamics, evolution, and mass function. The best current images resolve the near-infrared emission of the central parsec into thousands of stars with K-magnitudes up to 15 or 16 (see, e.g.,  Fig. 5-15). At this sensitivity, all red and most blue



■ Fig. 5-15

Very sharp 2 light-year \times 2 light-year view of the stars surrounding the supermassive black hole at the heart of the Milky Way, created with the European Southern Observatory's 8.2-m telescope atop Paranal, Chile. The colorization was produced by blending three images between 1.6 and 3.5 μ , using a color scheme in which *blue* is hot and *red* is cool. The location of Sagittarius A*, which coincides with the *center* of the Galaxy, is indicated by the two *yellow* arrows in the middle of the image (Photograph courtesy of R. Genzel et al. at the Max-Planck-Institut für Extraterrestrische Physik, and the European Southern Observatory)

supergiants, all red giants of spectral type later than K5, and all main sequence stars earlier than B2 are visible.

A group of more than two dozen stars (the IRS 16 complex) centered about $1''$ – $2''$ east of Sagittarius A* is quite prominent in this region. A second compact group of bright stars (the IRS 13 complex) lies $3.5''$ to the southwest, and the additional concentration of light within $1''$ of the middle marks the presence of the so-called Sagittarius A* cluster, a grouping of stars orbiting within 0.12 light-year of the supermassive black hole.

Attempts to determine the global properties of the stellar population close to the black hole are based on the premise that there exists an underlying dynamically relaxed distribution accounting for most of the stars (and most of the mass). The young and bright stars (such as those within the IRS 16 and IRS 13 complexes) may still be unrelaxed or are perhaps susceptible to other environmental factors, such as stellar collisions, and are presumably not reliable tracers. Instead, it is the surface density and surface brightness of the old, faint stars that apparently trace the equilibrium distribution.

At the galactic center, the luminosity profile is consistent with a stellar volume density that follows an r^{-2} power law (see, e.g., Sellgren et al. 1990; Serabyn and Morris 1996) from a projected radius $\sim 40''$ (≈ 100 pc) down to about $10''$ (≈ 0.4 pc). At least for radii > 3 pc, the stars cannot be directly influenced by the central black hole. Instead, a stellar distribution following

an r^{-2} power law is consistent with an *isothermal* profile, meaning that it may be characterized by a single energy variable equivalent to the temperature in a gaseous system.

The K-band spectra of the IRS 16/IRS 13 stars contain strong HeI lines, together with products (nitrogen and carbon) in their outer atmospheres of significant nucleosynthesis. These “HeI” stars therefore appear to be members of the blue supergiant variety, characterized by an initial mass $> 40 M_{\odot}$, that have evolved off the main sequence. They are probably on their way to becoming Wolf–Rayet stars and then supernovae.

Wolf–Rayet stars are hot ($\sim 25,000$ – $50,000$ K), massive ($> 20 M_{\odot}$) objects with a high rate of mass loss. Strong, broad emission lines, consistent with wind speeds in excess of 750 – $1,000$ km s^{-1} , are apparently produced in the material blown off their surface. But how these stars were assembled in the first place is not entirely clear, though they may represent the most massive members of a burst of star formation that occurred between two and nine million years ago. Presumably several hundred OB stars and thousands of others all formed at about the same time. Given their mass, there would have been sufficient time by now for the HeI stars to have evolved off the main sequence.

In contrast, the integrated spectrum of the Sagittarius A* cluster, concentrated within $1''$ of the radio source, is blue and featureless (Gezari et al. 2002), suggesting stars hotter than K-type giants. With K-band magnitudes of ~ 14 – 16 , these cluster members therefore appear to be early B or late O stars, and presumably quite young (< 20 Myr), so how could they have formed so close to the supermassive black hole, where the extreme conditions would have inhibited star formation? Gas near the galactic center would have to be compressed to densities five orders of magnitude higher than is typically found in the interstellar medium in order to overcome the strong magnetic fields (\sim mG), large turbulent velocities (~ 10 km s^{-1}), high temperatures, and strong tidal forces induced by the black hole, all of which conspire to prevent condensations of gas from forming protostars (Morris 1993).


Perhaps a clue to their formation is given by the fact that the starburst (early-type) stars exhibit a well-defined overall angular momentum unlike that expected on the basis of the general galactic rotation. Their line-of-sight velocities follow a rotation pattern with blue-shifted radial velocities north and red-shifted velocities south of the dynamical center. The corresponding rotation axis of this sample is aligned with the east-west direction. It therefore appears that the early-type stars are rotating clockwise (on the sky), counter to the overall galactic rotation. Their average rotation rate (~ 150 km s^{-1}) is relatively large, consistent with an enclosed central concentration of two to three million solar masses. By comparison, the late-type stars exhibit a slow rotation rate of only a few tens of km s^{-1} .


Thus, most of the early-type stars within $10''$ (≈ 0.4 pc) of Sagittarius A* appear to be unrelaxed. It is natural to wonder how they came to reside in this chaotic environment. Did they form in situ, or did they migrate toward the black hole from somewhere else? Given their short life span, it is difficult to see how they could have simply diffused to the center via two-body interactions. The relaxation time of stars within the cusp is greater than 10^8 years (Alexander 2003). By comparison, the lifetime of $\sim 10 M_{\odot}$ stars is about an order of magnitude shorter.

A single star would have insufficient time to sink gravitationally to the center via dynamical friction, but the timescale for this process goes as the inverse of the object’s mass. A compact young cluster could in principle migrate to the central parsec before getting tidally disrupted (Gerhard 2001), but detailed simulations of this process show that the cluster must be very massive ($\gg 10^4 M_{\odot}$) and very compact (< 0.2 – 0.4 pc) in order to spiral into the center within the lifetime of its O-stars (a few Myrs).


In summary, the stellar cusp centered on the massive black hole consists not only of old, low-mass stars with properties similar to those found in the galactic bulge, but it also contains a surprising number of unrelaxed, apparently young massive objects, including very short-lived HeI emission line stars with masses 30–100 M_{\odot} . The nucleus is evidently a region undergoing episodic star formation, producing an influx of early-type unrelaxed stars into the general mix.

6 The Environment Surrounding Sagittarius A*

At least some of the local hot plasma within a few parsecs of Sagittarius A* must be injected into the interstellar medium via stellar winds, and the diffuse X-ray emission seen in  Fig. 5-14 therefore constitutes an excellent probe of the gas dynamics near the black hole. Rather strong outflows exist in and around the nucleus, most of them due to the early-type stars discussed above. Measurements of high velocities associated with IR sources in Sgr A West (e.g., Krabbe et al. 1991) and in IRS 16 (Geballe et al. 1991), among others, provide clear evidence of a hypersonic wind, with a velocity $v_w \sim 500\text{--}1,000 \text{ km s}^{-1}$, a number density $n_w \sim 10^{3\text{--}4} \text{ cm}^{-3}$ near the mass-ejecting stars, and a total mass loss rate $\dot{M}_w \sim 3\text{--}4 \times 10^{-3} \dot{M}_{\odot}$, pervading the inner parsec of the Galaxy.

Comprehensive high-resolution numerical simulations of the wind–wind interactions using a detailed suite of stellar wind sources and their inferred wind velocities and outflow rates suggest that wind–wind collisions create a complex configuration of shocks that efficiently convert the kinetic energy of the outflows into internal energy of the gas (Rockefeller et al. 2004).  Figure 5-16 shows isosurfaces of specific internal energy in the central cubic parsec around 10,000 years after the beginning of the calculation. The dark surfaces indicate regions of gas with low specific internal energy; these tend to lie near the wind sources themselves. The gray surfaces mark regions of high specific internal energy, where gas has passed through multiple shocks. From simulations such as this, it is evident that about a quarter of the total energy in the central parsec is converted to internal energy via multiple shocks. The total kinetic energy of material there is $\sim 8 \times 10^{48}$ ergs, while the total internal energy is $\sim 3 \times 10^{48}$ ergs.

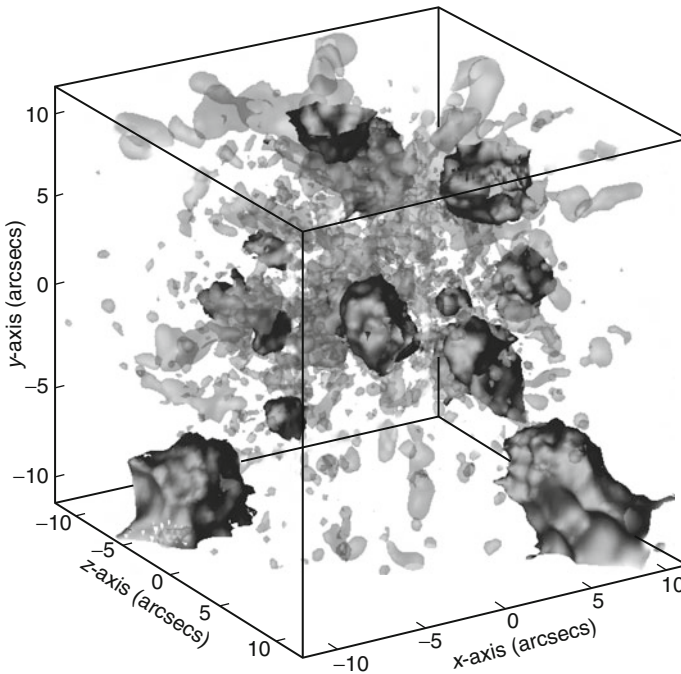
The environment that provides the plasma accreting onto the central black hole is therefore relatively clumpy and turbulent. This structure possibly explains why the Galaxy’s supermassive black hole is so underluminous compared to expectations, as discussed earlier in this chapter.

The rate at which Sagittarius A* ought to be accreting may be estimated using the classical Hoyle–Lyttleton theory (Hoyle and Lyttleton 1941), in which interstellar matter is captured by the central accretor once its total specific energy (gravitational potential energy per unit mass plus kinetic energy per unit mass) is negative. Of course, the Hoyle–Lyttleton accretion rate depends sensitively on the (as yet imprecisely known) value of v_w and n_w . The former may be as high as $1,000 \text{ km s}^{-1}$ if the dominant flow past the black hole is unshocked; it may be as low as one-fourth of this value otherwise. The density may be as high as $\sim 10^4 \text{ cm}^{-3}$ upstream of a shock, or it could be as low as the value 26 cm^{-3} , derived empirically from the *Chandra* observations of the diffuse X-ray emission (see  Fig. 5-14). In either case, taking the most conservative view, the accretion rate \dot{M} onto Sagittarius A* derived from this simple theory appears to be $> 10^{20} \text{ g s}^{-1}$. But this value is several orders of magnitude larger than that indicated by the observations.

However, in the absence of a uniform flow past the accretor, only those clumps with relatively low specific angular momentum descend deeply into the potential well, which may

partially reduce the overall accretion rate. In addition, the tenuous plasma is magnetized, which also means that magnetic energy must be liberated as the gas is compressed toward smaller radii. The importance of magnetic field dissipation was recognized from the earliest thinking on this subject (see, e.g., Shvartsman 1971; Melia 1994). The highly ionized plasma “freezes” the magnetic field and intensifies it due to flux conservation in the flow converging toward the black hole.

These ideas were put to a numerical test in 2002, with the first magnetohydrodynamic (MHD) simulation of Hoyle–Lyttleton accretion onto Sagittarius A* (Igumenshchev and Narayan 2002). Though several important caveats must be appended to this work, the results are quite promising in demonstrating how the entrained magnetic field may modify the dynamics of the flow to the point where the accretion rate onto Sagittarius A* is heavily attenuated. The magnetic field intensity grows with increasing compression of the gas toward smaller radii, eventually reaching superequipartition values, at which reconnection takes place. Converting magnetic energy into heat, this process produces buoyancy in the gas, inhibiting the infall rate.



■ Fig. 5-16

About 10,000 years after the winds are “turned on” in a three-dimensional hydrodynamics simulation, the 2–10 keV luminosity from the central 3 pc of the Galaxy reaches steady state. Shown here are the isosurfaces of specific internal energy of the shocked gas within the inner 20'' cube of the Galaxy. The line of sight is along the z-axis. The darkest surfaces correspond to a specific internal energy of 2.5×10^{12} ergs g^{-1} ; the gray surfaces correspond to 3.8×10^{15} ergs g^{-1} (Image from Rockefeller et al. 2004)

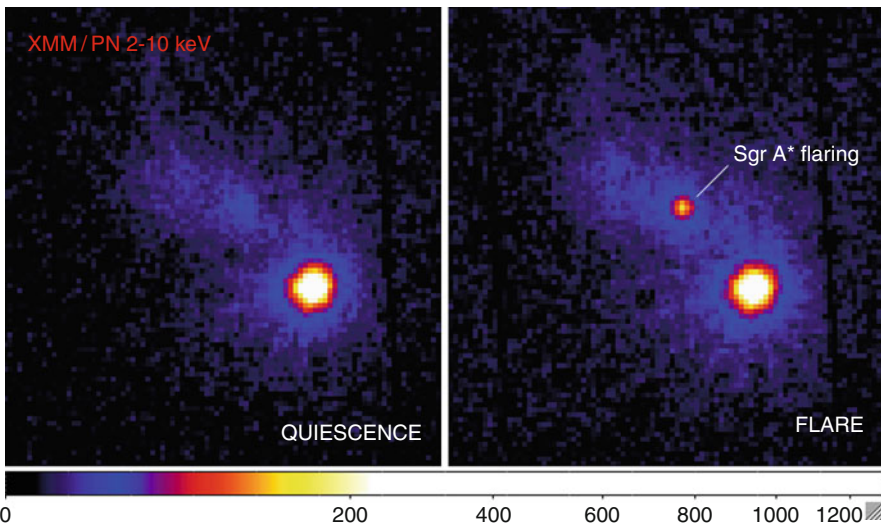
A large fraction of the captured gas actually attains escape energy and leaves the black hole system altogether. In the end, only a few percent of the inflowing plasma reaches the black hole's event horizon.

A disk may be present in Sagittarius A*, but it is not necessarily the sole contributor to its spectrum. Like many other radiating, compact objects, Sagittarius A* appears to manifest itself through several principal emitting regions; the challenge is to determine how all of these function together to produce the radiative flux and polarization fraction seen originating from the galactic nucleus.

The infalling matter appears to be radiatively inefficient, at least until it settles into a disk. Of course, this is not obvious, since the captured gas is highly ionized (and magnetized), so it radiates via several mechanisms, including bremsstrahlung, cyclosynchrotron (i.e., the full range of cyclotron to synchrotron emission in a magnetic field), and inverse Compton scattering. Thus, the overall power depends on several critical physical parameters: the particle density, the temperature (or Lorentz factor, if the dominant emitters are nonthermal), and the magnetic field.

But Sagittarius A* is not a typical AGN. The rate at which it accretes is apparently less than $\sim 10^{19} \text{ g s}^{-1}$, and a naive integration of the particle number density implied by this value from, say, 40 Schwarzschild radii out to infinity yields a scattering optical depth smaller than 10^{-16} . Clearly, the medium surrounding the black hole is extremely thin, and the integrated emissivity must therefore be correspondingly small.

This can change dramatically once the infalling gas circularizes and forms a disk. The various hydrodynamic and magnetohydrodynamic simulations discussed above indicate that the accreted specific angular momentum Λ (in units of cr_S , where r_S is the Schwarzschild radius) can vary by 50% over < 200 years, with an average equilibrium value of $\sim 40 \pm 10$. However, Λ



■ Fig. 5-17

Left: X-ray image (approximately 10 pc on each side) of the quiescent phase during the 50 min prior to the flare. *Right:* A 50-min integrated image during the flare (Image from Trap et al. 2010)

is never zero, meaning that the gas cannot simply flow radially inward all the way to the event horizon. Nonetheless, even with a possibly large quantity of angular momentum present in the environment surrounding the nucleus, relatively little specific angular momentum is accreted due to the clumpy nature of the environment (see [Fig. 5-16](#)). The associated variability in the sign of the components of Λ suggests that if an accretion disk forms at all, it dissolves and reforms (perhaps) with an opposite sense of spin on a timescale of around 100 years or less.

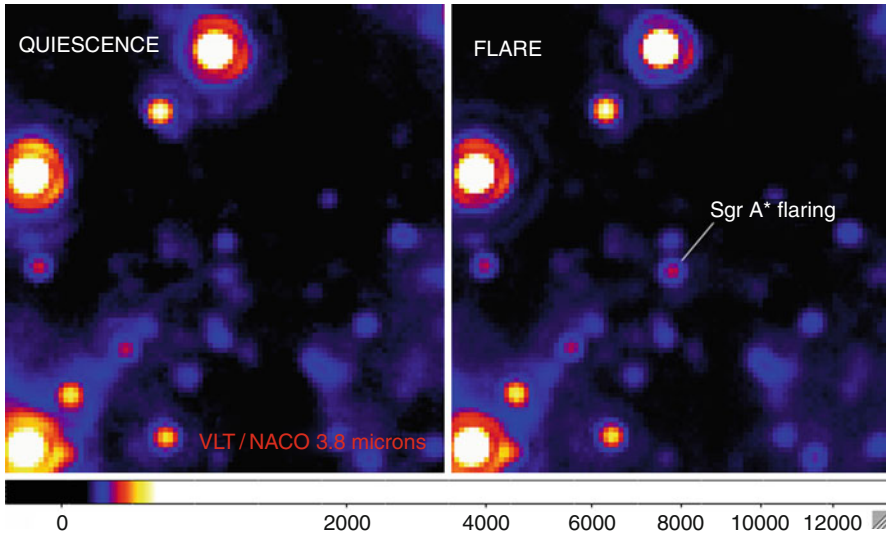
Perhaps not surprisingly, most of the information pertaining to Sagittarius A* has been gleaned (not during its quiescent phase, but rather) from an examination of the radiative emission associated with its flaring state. About once per day, Sagittarius A*'s X-ray luminosity increases by a factor of up to several hundred, lasting a few hours. A ~ 10 -min substructure in the lightcurve of the eruption implies (via light-travel arguments) that these events occur within 10–15 Schwarzschild radii of the event horizon. [Figure 5-17](#) shows an image of such a flare obtained with XMM-Newton in 2007 (Trap et al. 2010). Infrared flares occur more frequently, around four per day, and have been observed in many passbands. An example of such an event, contemporaneous with that presented in [Fig. 5-17](#), is shown in [Fig. 5-18](#). The origin of these flares is still unclear, though the infrared photons are almost certainly produced by synchrotron processes, whereas the X-rays are presumably due to inverse Compton scattering.

This pattern of flaring challenges disruption mechanisms of the accretion flow as the origin of the outbursts, since they rely on a temporary storage of mass and energy. This energy should indeed be released at once during the event, with a radiation efficiency of a few percent. But the weak (average) accretion rate of the black hole seems insufficient to accumulate the required energy on such short timescales. Instead, the flares may be due to the stochastic infall and tidal disruption of the gas clumps illustrated in [Fig. 5-16](#) (see also Tagger and Melia 2006), or perhaps even to small bodies such as asteroids or comets (see, e.g., Cadez et al. 2008).

7 Strong Field Physics

Though a black hole's radiative characteristics are produced by proxy, via the compression and emissivity of matter accreting toward it or expelled into a surrounding nonthermal halo or jet, the emitting plasma in Sagittarius A* is sufficiently close to its event horizon that the signature of strong gravity ought to emerge in any image made of this region. The mm/sub-mm radiation originates from only a handful of Schwarzschild radii above the event horizon (see, e.g., [Fig. 5-12](#)), so it must be subject to significant light bending and area amplification that, given suitable conditions, can lead to a shadow observable at Earth's distance from the galactic center. A *spinning* black hole is expected to produce even stronger effects, including distortions to the shadow, from which the spin itself might be measured. And with spin, gravity would acquire a dependence on polar angle that can sometimes induce a precession in the disk.

The radio waves at ~ 1 mm in Sagittarius A* are emitted within a region roughly the size of Earth's orbit about the Sun, corresponding to ~ 15 Schwarzschild radii for a black hole with a mass of $3\text{--}4 \times 10^6 M_{\odot}$. The synchrotron emission at mm/sub-mm wavelengths, moreover, appears to be optically thin, so the medium surrounding the black hole is transparent at these



■ Fig. 5-18

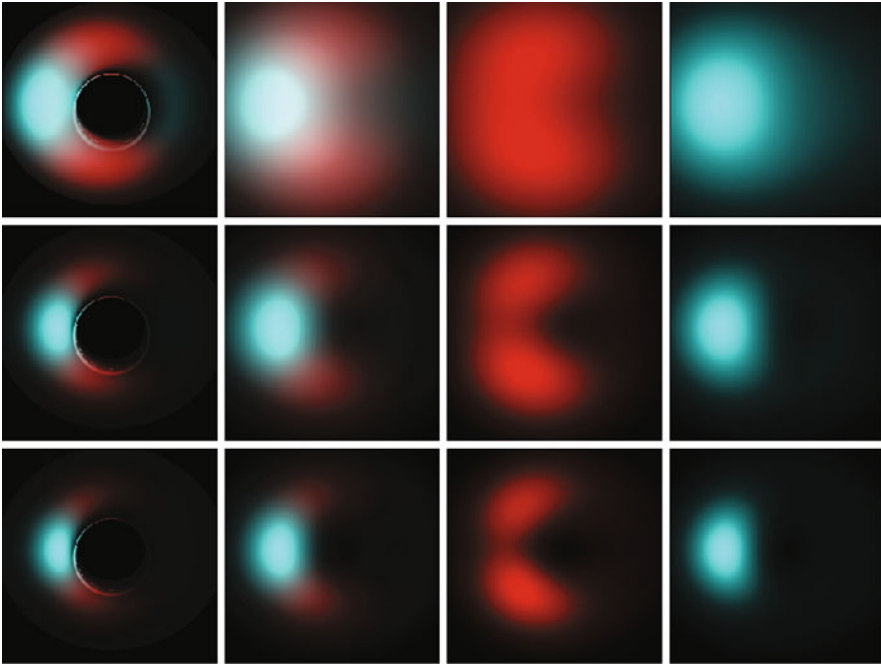
Left: A near-infrared image (approximately 1 pc on each side) of the quiescent phase during the 6 min prior to the flare. *Right:* A 6-min integrated image near the peak of the flare (Image from Trap et al. 2010)

wavelengths. It should therefore be possible to see the effects of light-bending and absorption in this portion of the spectrum.

The idea of a black hole creating a dark depression in an otherwise brightly lit background of emission was raised long ago by Bardeen (1974), who described the idealized appearance of such an object in front of a planar-emitting source. The intensity from the accretion disk at infinity may be obtained with the simplifying assumption that the emitter is geometrically thin, with a uniform slab structure in the vertical direction. Other than this approximation, it is straightforward to incorporate all of the relativistic effects into the calculation, including frame dragging, gravitational redshift, light bending, and Doppler boosting (see, e.g., Bromley et al. 1997).

This ray-tracing calculation produces a pixelized image, with specific intensities at the detector. Fluxes are then calculated by summing over the pixels in the image, taking into account the physical size of the detector array and the distance to the source. The ray-tracing method used to produce the images shown in ► Fig. 5-19 is described more fully in Falcke et al. (2000) and Bromley et al. (2001).

But to properly simulate an *observed* image, two additional effects must be taken into account: interstellar scattering and the finite telescope resolution achievable from the ground. The first of these may be incorporated by smoothing the image with an appropriately scaled elliptical Gaussian, based on the measured properties of the interstellar medium along the line-of-sight. The effects of a finite telescope resolution may then be added in an idealized form by convolving the smoothed image with a spherical Gaussian point-spread function, whose width



■ Fig. 5-19

Computer simulation of Sagittarius A*'s "shadow," created by its strong gravitational field, which bends light incident from the rear to create the darkness, seen at three different wavelengths. The *top* row shows emission at 1.5 mm, the *middle* row is at 1 mm, and the *bottom* row corresponds to 0.67 mm. The images in the first column show the disk as it would appear without the blurring effects due to propagation of the radiowaves through the interstellar medium. The second column shows the processed images when the finite very large baseline interferometry resolution and interstellar scattering are taken into account. The two rightmost columns give the vertical and horizontal components of the polarized emission. Throughout, red pixels designate vertically polarized light, and cyan corresponds to horizontal polarization. The pixel brightness in all images scales linearly with flux (Image from Bromley et al. 2001)

reflects the possible resolution of a global interferometer with 8,000 km baselines (Krichbaum 1996).

The illustrative panels shown in ● Fig. 5-19 demonstrate the feasibility of imaging Sagittarius A* polarimetrically at mm/sub-mm wavelengths. The size of the shadow, roughly 5 Schwarzschild radii in diameter, represents a projected size of $\sim 40 \mu\text{arcsec}$, which is already within a factor of two to three times the current capability. The technical methods to achieve the necessary resolution to form these images directly are currently being developed for wavelengths shortward of ~ 1.3 mm. Depending on how short a wavelength is required, the projected timescale for completing the telescope array may be less than ten years from the time of this writing. Finally, the accretion flow appears to be optically thin to electron scattering at X-ray

wavelengths as well, so in principle, the shadow may also be detectable with space-based X-ray interferometry.

Cross-References

- [Astrophysics of Galactic Charged Cosmic Rays](#)
- [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- [Galactic Neutral Hydrogen](#)
- [Gamma-Ray Emission of Supernova Remnants and the Origin of Galactic Cosmic Rays](#)
- [High-Velocity Clouds](#)
- [History of DarkMatter in Galaxies](#)
- [Interstellar PAHs and Dust](#)
- [Magnetic Fields in Galaxies](#)
- [Open Clusters and their Role in the Galaxy](#)
- [The Infrared Galaxy](#)

References

- Alexander, T. 2003, in *The Galactic Black Hole: Lectures on General Relativity and Astrophysics* (Bristol, UK: The Institute of Physics Publishing), 246
- Allen, D. A., Hyland, A. R., & Hillier, D. J. 1990, *MNRAS*, 244, 706
- Anantharamaiah, K. R., Pedlar, A., Ekers, R. D., & Goss, W. M. 1991, *MNRAS*, 249, 262
- Backer, D. C. 1988, in *AIP Conference Proceedings*, 174, (New York: AIP), 111
- Baganoff, F. K., Maeda, Y., Morris, M., Bautz, M. W., Brandt, W. N., Cui, W., Doty, J. P. et al. 2003, *ApJ*, 591, 891
- Balick, B., & Brown, R. L. 1974, *ApJ*, 194, 265
- Bardeen, J. M. 1974, in *Gravitational Radiation and Gravitational Collapse*, (Dordrecht: D. Reidel Publishing Company), 132
- Becklin, E. E., Gatley, I., & Werner, M. W. 1982, *ApJ*, 258, 135
- Blaauw, A., Gum, C. S., Pawsey, J. L., & Westerhout, G. 1960, *MNRAS*, 121, 123
- Bower, G. C., & Backer, D. C. 1998, *ApJ*, 496, L97
- Bower, G. C., Goss, W. M., Falcke, H., Backer, D. C., & Lithwick, Y. 2006, *ApJ*, 648, L127
- Bromley, B., Chen, K., & Miller, W. A. 1997, *ApJ*, 475, 57
- Bromley, B., Melia, F., & Liu, S. 2001, *ApJ*, 555, L83
- Cadez, A., Calvani, M., & Kostic, U. 2008, *A&A*, 487, 527
- Davidson, J. A., Werner, M. W., Wu, X., Lester, D. F., Harvey, P. M., Joy, M., & Morris, M. 1992, *ApJ*, 387, 189
- Eckart, A., & Genzel, R. 1996, *Nature*, 383, 415
- Ekers, R. D., van Gorkom, J. H., Schwarz, U. J., & Goss, W. M. 1983, *A&A*, 122, 143
- Falcke, H., Markoff, S., & Bower, G. C. 2009, *A&A*, 496, 77
- Falcke, H., Melia, F., & Agol, E. 2000, *ApJ*, 528, L13
- Fromerth, M. J., Melia, F., & Leahy, D. A. 2001, *ApJ*, 547, 129
- Geballe, T. R., Krisciunas, K., Bailey, J. A., & Wade, R. 1991, *ApJ*, 370, L73
- Geballe, T. R., Wade, R., Krisciunas, K., Gatley, I., & Bird, M. C. 1987, *ApJ*, 320, 562
- Genzel, R., Thatte, N., Krabbe, A., Kroker, H., & Tacconi-Garman, L. E. 1996, *ApJ*, 472, 153
- Gerhard, O. 2001, *ApJ*, 546, L39
- Gezari, S., Ghez, A. M., Becklin, E. E. et al. 2002, *ApJ*, 576, 790
- Ghez, A., Duchene, G., Matthews, K., Hornstein, S. D., Tanner, A., Larkin, J., Morris, M. et al. 2003, *ApJ*, 586, L127
- Ghez, A., Klein, B. L., Morris, M., & Becklin, E. E. 1998, *ApJ*, 509, 678
- Gusten, R., Genzel, R., Wright, M. C. H., Jaffe, D. T., Stutzki, J., & Harris, A. I. 1987, *ApJ*, 318, 124
- Gwinn, C. R., Danen, R. M., Tran, T. K., Middleditch, J., & Ozernoy, L. M. 1991, *ApJ*, 381, L43

- Hall, D. N. C., Kleinmann, S. G., & Scoville, N. Z. 1982, *ApJ*, 260, L53
- Haller, J. W., Rieke, M. J., Rieke, G. H., Tamblyn, P., Close, L., & Melia, F. 1996, *ApJ*, 456, 194
- Hornstein, S. D. et al. 2007, *ApJ*, 667, 900
- Hoyle, F., & Lyttleton, R. A. 1941, *MNRAS*, 101, 227
- Igumenshchev, I. V., & Narayan, R. 2002, *ApJ*, 566, 137
- Jackson, J. M., Geis, N., Genzel, R., Harris, A. I., Madden, S., Poglitsch, A., Stacey, G. J., Townes, C. H. 1993, *ApJ*, 402, 173
- Koyama, K., Maeda, Y., Sonobe, T., Takeshima, T., Tanaka, Y., & Yamauchi, S. 1996, *PASJ*, 48, 249
- Krabbe, A., Genzel, R., Drapatz, S., & Rotaciuc, V. 1991, *ApJ*, 382, L19
- Krichbaum, T. P. 1996, in *Science with Large Millimeter Arrays* (Berlin: Springer), 95
- LaRosa, T. N., Kassim, N. E., & Lazio, T. J. W. 2000, *AJ*, 119, 207
- Liu, S., & Melia, F. 2001, *ApJ*, 561, L77
- Lo, K. Y., Cohen, M. H., Readhead, A. S. C., & Backer, D. C. 1981, *ApJ*, 249, 504
- Lo, K. Y., & Claussen, M. J. 1983, *Nature*, 306, 647
- Lo, K. Y., Schilizzi, R. T., Cohen, M. H., & Ross, H. N. 1985, *ApJ*, 202, L63
- Lo, K. Y., Shen, Z.-Q., Zhao, J.-H., & Ho, P. T. P. 1998, *ApJ*, 508, L61
- Lynden-Bell, D., & Rees, M. J. 1971, *MNRAS*, 152, 461
- Maeda, Y., Baganoff, F. K., Feigelson, E. D., Morris, M., Bautz, M. W., Brandt, W. N., & Burrows D. N. et al. 2002, *ApJ*, 570, 671
- McGinn, M. T., Sellgren, K., Becklin, E. E., & Hall, D. N. B. 1989, *ApJ*, 338, 824
- Melia, F. 2003, *The Black Hole at the Center of Our Galaxy* (Princeton: Princeton University Press)
- Melia, F. 1994, *ApJ*, 426, 577
- Morris, M. 1993, *ApJ*, 408, 496
- Murakami, H., Koyama, K., & Maeda, Y. 2001, *ApJ*, 558, 687
- Pedlar, A., Anantharamaiah, K. R., Ekers, R. D., Goss, W. M., van Gorkom, J. H., Schwarz, U. J., & Zhao, J.-H. 1989, *ApJ*, 342, 769
- Quataert, E. 2002, *ApJ*, 575, 855
- Rieke, G. H., & Rieke, M. J. 1989, *ApJ*, 344, L5
- Rockefeller, G., Fryer, C. L., Melia, F. & Warren, M. S. 2004, *ApJ*, 604, 662
- Romani, R., Narayan, R., & Blandford, R. 1986, *MNRAS*, 220, 19
- Sanders, R. H., & Prendergast, K. G. 1974, *ApJ*, 188, 489
- Schodel, R., Ott, T., Genzel, R., Hofmann, R., Lehnert, M., Eckart, A., & Mouawad, N. et al. 2002, *Nature*, 419, 694
- Sellgren, K., McGinn, M. T., Becklin, E. E., & Hall, D. N. 1990, *ApJ*, 359, 112
- Serabyn, E., & Morris, M. 1996, *Nature*, 382, 602
- Shvartsman, V. F. 1971, *Soviet Astronomy*, 15, 37
- Snowden, S. L., Egger, R., Freyberg, M. J., McCammon, D., Plucinsky, P. P., Sanders, W. T., Schmitt, J. H. M., Truemper, J., & Voges, W. 1997, *ApJ*, 485, 125
- Tagger, M., & Melia, F. 2006, *ApJ*, 636, L33
- Tanaka, Y., Koyama, K., Maeda, Y., & Sonobe, T. 2000, *PASJ*, 52, L25
- Trap, G. et al. 2010, *Adv Space Res*, 45, 507
- Wang, Q. D., Gotthelf, E. V., & Lang, C. C. 2002, *Nature*, 415, 148
- Yusef-Zadeh, F., & Morris, M. 1987, *ApJ*, 320, 545
- Yusef-Zadeh, F., Morris, M., & Chance, D. 1984, *Nature*, 310, 557

6 The Galactic Bulge

R. Michael Rich

Department of Physics and Astronomy, University of California,
Los Angeles, CA, USA

1	<i>Introduction</i>	273
1.1	Overview, Scope, and Definition	276
1.2	A Brief History	278
2	<i>The Age and Population of the Galactic Bulge</i>	283
2.1	Evidence for Minority Populations of Intermediate and Younger Age	286
2.2	Microlensed Dwarfs: A Young, Metal-Rich Population?	290
2.3	The Luminosity Function	293
2.4	Globular Clusters	294
3	<i>Composition</i>	295
3.1	Optical Spectroscopy	296
3.2	Infrared Spectroscopy	298
3.3	Composition and Comparison with Other Populations	300
3.4	Na and Al	302
3.5	Heavy Elements	304
4	<i>Kinematics</i>	306
4.1	Stellar Radial Velocity Surveys	309
4.2	Proper-Motion Studies	314
5	<i>Kinematics and Composition</i>	317
5.1	Are There Subcomponents in the Bulge Abundance Distribution?	317
6	<i>Structure</i>	319
6.1	The X-Shaped Bulge	324
6.2	A Classical Bulge?	327
7	<i>The Milky Way Bulge in an Extragalactic Context</i>	328
8	<i>Theories for the Formation of the Bulge</i>	334
9	<i>Future Surveys</i>	335
9.1	Ground-Based Imaging Surveys	335
9.2	Spectroscopic Surveys	337

9.3	Radio Surveys	338
9.4	Space-Based Surveys	338
10	<i>Observational Challenges for the Future</i>	339
	<i>References</i>	341

Abstract: The central bulge of the galaxy is considered from the standpoint of stellar populations, ages, composition, kinematics, structure, and extragalactic context. The central bulge is a 3:1 bar viewed edge-on mass of $2 \times 10^{10} M_{\odot}$; the major axis is oriented $\sim 20^{\circ}$ toward the first quadrant. At $|b| > 8^{\circ}$ much of the mass appears to be in an X-shaped distribution that resembles similar structures seen in extragalactic boxy bars. The bar exhibits cylindrical rotation and has all the hallmarks of structures that evolve dynamically from massive disks. Although there is a compelling case for secular evolution, the evidence from the color-magnitude diagram is that the bulge is ~ 10 Gyr old, with no significant difference between the age of the bulge and the metal-rich globular clusters. Near the Galactic nucleus, and to a lesser extent, within the inner 100–200 pc, there is evidence of ongoing star formation and intermediate-age stars. A “long bar” oriented at $\sim 45^{\circ}$ and with vertical thickness comparable to that of the disk may be a separate structure or part of the main bar, and a nuclear disk or bar may be present. A young and intermediate-age stellar population is confined to the central 100 pc, predominantly toward the nucleus. The kinematics are also consistent with most of the mass in the bulge being in the bar, with $< 10\%$ of the mass being in a “classical” bulge.

The bulge outside of the inner 200 pc is old, globular cluster-aged, with $[\text{Fe}/\text{H}] \sim \text{solar}$, with a range spanning -1.5 – -0.5 dex in $[\text{Fe}/\text{H}]$ and a gradient of $[\text{Fe}/\text{H}] = -0.6 \text{ dex kpc}^{-1}$ at $|b| > 4^{\circ}$; there is at present no evidence for a gradient in $[\text{Fe}/\text{H}]$ or $[\alpha/\text{Fe}]$ for $|b| < 4^{\circ}$. The alpha elements are enhanced over the full Galactic bulge; models of chemical evolution associate the alpha enhancement and iron abundance distribution with a rapid (< 1 Gyr) formation timescale. This is consistent with trends observed for heavy elements, which are consistent with a pure r-process enrichment pattern. Although there are hints that the bulge system has subpopulations that can be distinguished based on chemodynamical properties, the reality of population components remains a matter of debate.

The luminous mass of the bulge is not a spheroidal stellar system comparable to bright elliptical galaxies; rather, it meets the criteria to be classified as a pseudobulge according to Kormendy and Kennicutt (2004). While overwhelmingly an old stellar population, the kinematics and structure of the bar are consistent with it having evolved secularly from a massive disk.

The review concludes with an examination of the major new ground- and space-based surveys of the bulge that will be carried out in the time frame 2015–2025 and a recital of observational challenges for the next decade.


Keywords: TK

1 Introduction

We must conclude, then, that in the central region of the Andromeda Nebula we have a metal poor Population II, which reaches -3^m for the brightest stars, and that underlying it there is a very much denser sheet of old stars, probably something like those in M67 or NGC 6752. We can be certain that these are enriched stars, because the cyanogen bands are strong, and so the metal/hydrogen ratio is very much closer to what we observe in the sun and in the present interstellar medium than to what is observed for Population II. And the process of enrichment has taken very little time. After the first generation of stars has been formed, we can hardly speak of a “generation”, because the enrichment takes place so soon, and there is probably very little time difference. So the CN giants

that contribute most of the light in the nuclear region of the Nebula must also be called old stars; they are not young. – W. Baade in *The Evolution of Galaxies and Stellar Populations* p. 256 (1961)





Examples of all of the major stellar populations may be found within a reasonable vicinity of the Sun; the nearest globular clusters are just over 3 kpc distant, while the nearest disk and halo dwarfs can be found within 100 pc. The Galactic bulge is another matter. For the most part, the bulge population is 8 kpc distant, and over much of its area is obscured by tens of magnitudes of visual extinction. The nearest comparable population is the bulge of M31, one hundred times more distant. Until the last two decades, studies of the M31 bulge that population were limited to photometry and spectroscopy of its integrated light. These factors meant that much more time was required for an understanding of the bulge population to be achieved, compared with the disk, halo, and globular clusters. These challenges also meant that the bulge subject area has historically experienced a greater number of wrong turns and controversies; lingering issues, like the nature of the long bar, remain to the present day.

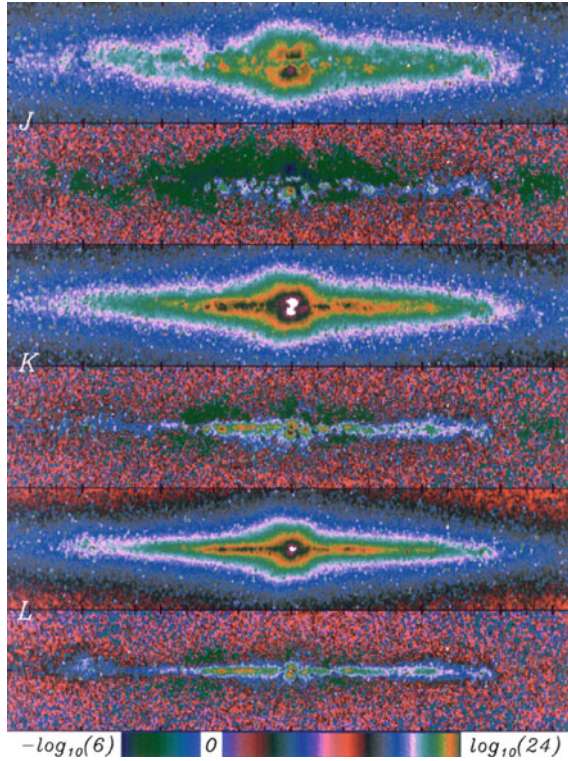
Considering that more than half of the light in the local universe is found in spheroids (Fukugita et al. 1998) with the Galactic bulge being the only example with individual stars readily available for study, the importance of this population can hardly be overstated.  [Figure 6-1](#) shows how dramatic the effect of reddening is on the bulge, even covering the relatively red wavelengths of the *DIRBE* instrument (Weiland et al. 1994). In this chapter, I will refer to the central region of the Milky Way galaxy as the “bulge” even though there is now growing evidence that most of the mass is in a bar structure.

We open with Baade’s prescient words because they define the essence of our present-day concept of the bulge: a roughly solar-metallicity stellar population that is as old as the halo globular clusters and exhibits chemical signatures of rapid formation. Of course, Baade’s conclusions were largely based on intuition rather than observations; our present-day picture is grounded on far more sturdy observational evidence.

The Galactic bulge and bulges of galaxies have been the subject of two IAU symposia, IAU 153 (1993) and IAU 245 (2008), and have been reviewed in numerous other IAU symposia concerning the Milky Way and external galaxies (e.g., Binney 2009). On three occasions, the bulge or bulges have been specifically reviewed in *Annual Reviews of Astronomy and Astrophysics* (Frogel 1988; Wyse et al. 1997; Kormendy and Kennicutt 2004). The structure of the bar and orbit theory is not reviewed here in detail; the interested reader may consult reviews by Sellwood and Wilkinson (1993), Gerhard (2002, 2011), Athanassoula (2008, 2012), Combes (2009), and Sellwood (2012; this volume) on that subject. We also do not review star count/population/kinematic models of the Milky Way, which are becoming increasingly capable in modeling the observed population of the bulge and are giving constraints on the bar structure and orientation. The principal efforts are the Besancon galaxy model (Robin et al. 2012) and the TRILEGAL model (Vanhollebeke et al. (2009); the latter article gives a useful compilation of derived properties for the bar).

The most significant recent observational developments include the completion of the first major surveys at low and high spectroscopic resolution and the completion of the Vista Variables in the Via Lactea (VVV) infrared imaging survey. In the next decade, the subject will change dramatically as spectroscopic samples of 10,000–100,000 stars are attained, UV/optical imaging surveys are merged with the VVV, and astrometric surveys from GAIA and other sources arrive.

This chapter is organized such that after this introductory section,  [Sect. 2](#) addresses the age of the bulge;  [Sect. 3](#) considers the composition;  [Sect. 4](#), the kinematics; and  [Sect. 5](#),

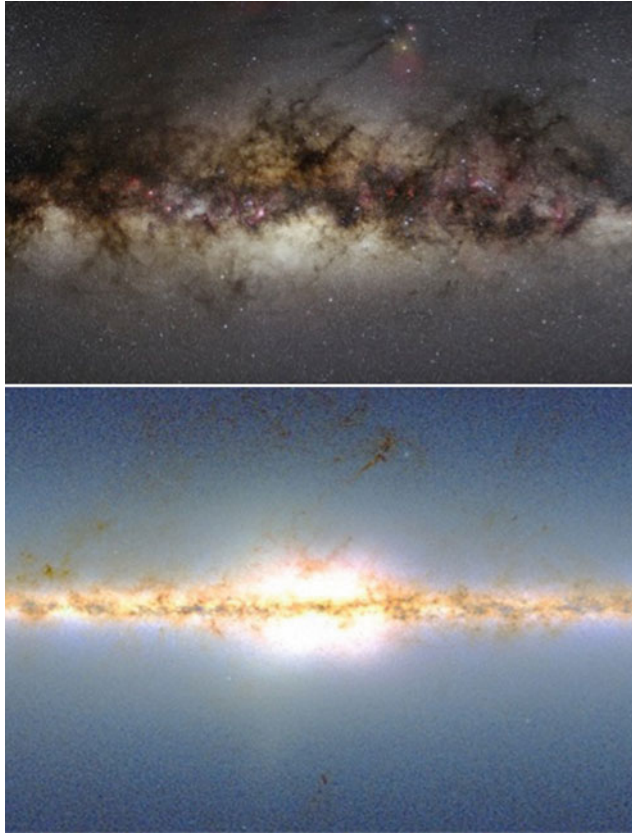


■ Fig. 6-1

J-, K-, and L-band surface brightnesses of the Galactic bulge, observed with the *DIRBE* instrument on the *COBE* satellite (Freudenreich (1998); see also Weiland et al. 1994; Dwek et al. 1995; Arendt et al. 1994). The *lowest panel* is the residuals resulting from the subtraction of a composite disk/bulge/reddening model from the data. Notice that the asymmetry of the bulge (*thick bar*) and disk (*long bar*) is clearly evident in the K and L bands; heavy interstellar reddening affects the J band (1.2 μm). The bulge asymmetry is due to the bar's alignment with the long axis aligned into the first quadrant $\sim 20^\circ$ off the minor axis (see [Sect. 6](#)). The range covers $l < 110^\circ$ and $b < 15^\circ$; *tick marks* are spaced at 10° in longitude and 3° in latitude

correlations between composition and kinematics, and other subpopulations. [Sect. 6](#) addresses the structure of the bulge, while [Sect. 7](#) places our bulge in an extragalactic context (note that Kormendy and Kennicutt (2004) consider in depth the issue of bulges and secular evolution). [Sect. 8](#) is a brief consideration of some theoretical issues. [Sect. 9](#) considers planned ground- and space-based surveys, while [Sect. 10](#) addresses some issues that remain as challenges to be addressed with the next generation of observations.

The region of the Galaxy that we refer to as the “bulge” appears to be predominantly what is now known to be a dynamical bar population. For the purpose of this review, I will emphasize the inner 1.5 kpc of the Galaxy, but will consider evidence for the spheroid at higher latitude. The total luminosity is $L_{\text{bol}} \sim 10^{10} L_\odot$ and the mass is $1\text{--}2 \times 10^{10} M_\odot$. While the term “bulge” is therefore almost certainly not correct, I will use the term throughout this review to mean this inner region of the Milky Way, which almost certainly includes more than the *COBE* image of the bulge in [Figs. 6-1](#) and [6-2](#). We should be mindful that the stellar halo almost certainly



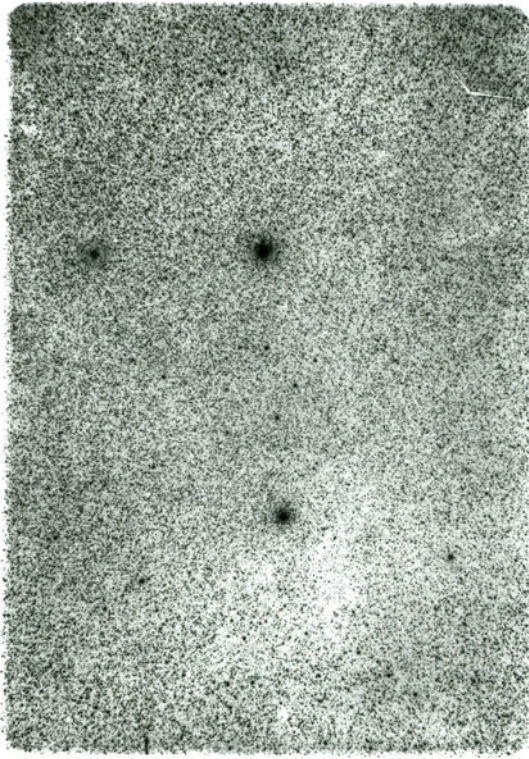
■ Fig. 6-2

The bulge from the Photopic Sky Survey (optical image) and 2MASS All-Sky Data (lower image). The Sagittarius dwarf spheroidal galaxy is faintly visible as a stream to the South of the bulge. The images span roughly $70^\circ \times 50^\circ$. Notice how the true form of the bulge is difficult to discern in the optical image, due to the heavy reddening the regions of greatest infrared surface brightness are obscured in the optical image

interpenetrates the bar, along with some component of the thick disk. At the Galactic center, there is well-documented evidence for a young stellar population, extending to the vicinity of the Galactic nucleus (e.g., Ghez & Morris 2009; Morris and Serabyn 1996; Ghez et al. 2003; Eisenhauer et al. 2005). This review is concerned with the stellar component of the bulge, and discussions of the cold gas and hot gas components are beyond the scope of this chapter. The bulge is known to be dominated by a hot, rotating (Rich et al. 2007a; Kunder et al. 2012) stellar population, with approximately solar metallicity (e.g., Rich 1988; McWilliam and Rich 1994) and globular cluster age (Ortolani et al. 1995; Clarkson et al. 2008).

1.1 Overview, Scope, and Definition

While a stellar population can be defined by ages, abundance, kinematics, and spatial distribution, the reality is that these variables are all interrelated. We therefore treat the observations



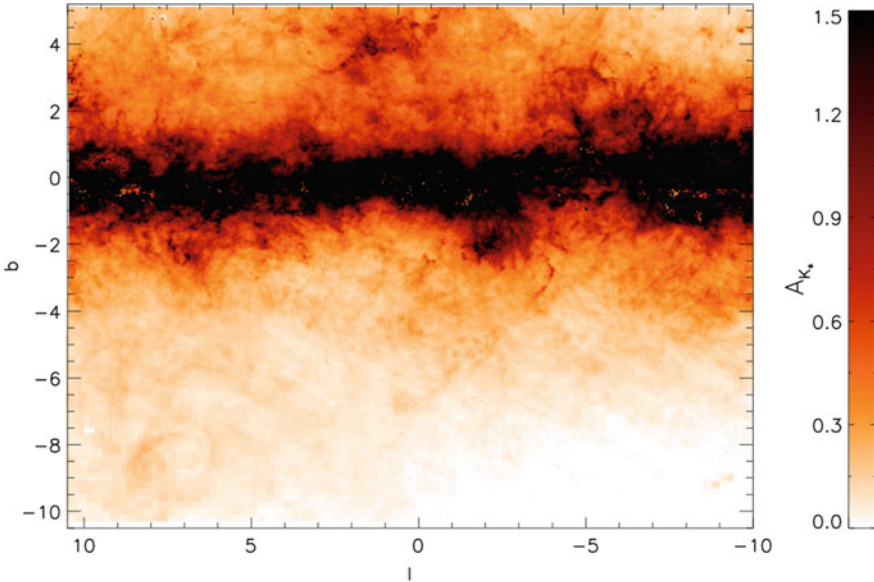
■ Fig. 6-3

Baade's Window, original plate by Baade. North up, east left. NGC 6528 (*lower cluster*) and NGC 6522 (*upper*). The field near NGC 6522 was favored because of the presence of the cluster as a kind of "standard" population, as well as the relative uniformity of extinction; however, much of the bulge has extinction lower than is found in this field. Photographic reproduction courtesy of S. Majewski, Carnegie Observatories

first and then consider more complex issues, such as correlations between abundances and kinematics and, finally, models of the formation of the bulge.

One frequently encounters terminology for specific bulge fields. This dates back to the time when observations were seriously affected by extinction and calibration of the reddening was complicated. Much of the bulge with $b < -4^\circ$ has $E(B - V) < 0.5$, but prior to the era of infrared surveys, it was very difficult to derive a quantitative reddening determination in a given field. Baade's Window ($l, b = 0.9^\circ, -3^\circ.9$) is centered on the globular cluster NGC 6522 and was designated in part because the region has relatively low and uniform extinction (► Fig. 6-3). Further, the metal-poor globular cluster NGC 6522 was used as a "standard source" to calibrate the reddening toward that field, which included the RR Lyrae stars used for measuring the distance to the center of the Galaxy, R_0 (Baade 1951).

Because of the intensive efforts devoted to the study of that region, Baade's Window has consequently been the subject of other studies: Arp's (1965) color-magnitude diagram (► Fig. 6-5) and the Blanco et al. (1984) grism (low-dispersion slitless spectroscopy) study of M giants. The Arp identifications were utilized by Rich (1988, 1990), McWilliam and Rich (1994)



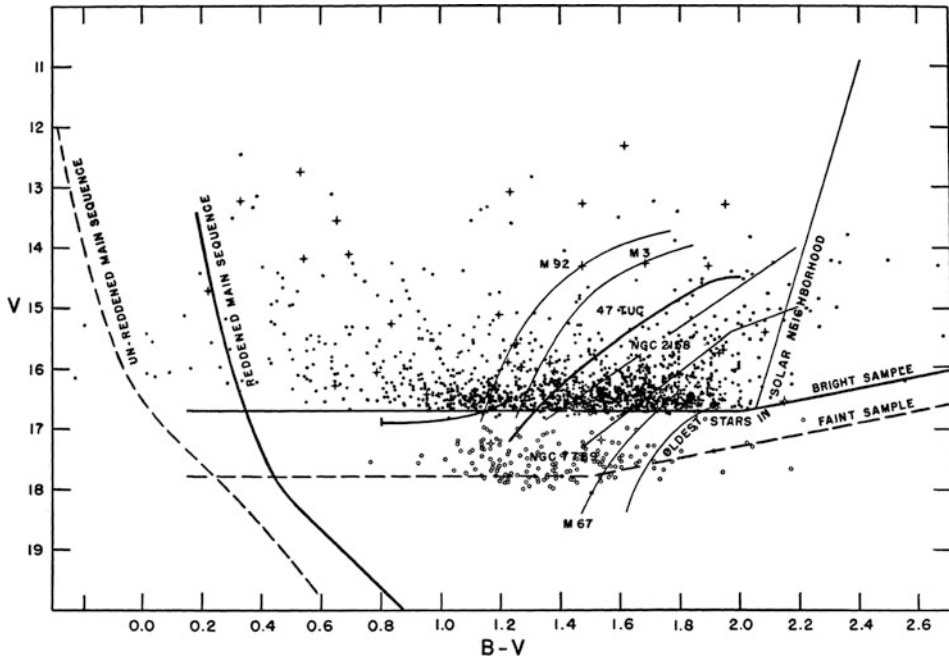
■ Fig. 6-4

K-band reddening map of the bulge from the infrared VVV survey (Gonzalez et al. 2012). l and b are in Galactic longitude and latitude in degrees

and Fulbright et al. (2006, 2007); and many subsequent studies (e.g., Ryde et al. 2009). The *Arp annulus* is the 1–4′ (the hash should be the arcmin designation) annular zone around NGC 6522, from which stars were measured from photographic photometry by Arp (1965). Other noteworthy fields include Plaut’s Field ($0, -8^\circ$), the subject of the van den Bergh and Herbst (1974) and other studies, and the lower-latitude Sgr I and II fields. The Sgr I field ($1^\circ.3, -2^\circ.65$) was the target of HST investigations (Holtzman et al. 1998). The Sagittarius dwarf spheroidal galaxy (Ibata et al. 1994) is ~ 10 kpc beyond the bulge, $(l, b) = 5^\circ.5, -14^\circ$, and was discovered in the course of a survey of the bulge (Ibata and Gilmore 1995) (● Fig. 6-4).

1.2 A Brief History

Victor Blanco once commented that, had astronomy originated in the Southern hemisphere, we would never have questioned where the Galactic center is, as it is so plainly visible when overhead. All-sky images clearly show a bulge structure toward Sagittarius, and it is surprising that there ever was controversy over the location of the Galactic center. The first quantitative work on the Sun’s position in the Milky Way was that of Shapley (1919) who used cluster variables along with the large number of bulge globular clusters, to show that the Galactic center lies in the direction of Sagittarius. The next major advance would be Baade’s work on stellar populations and his discovery of field RR Lyrae stars in the bulge from Mt. Wilson (Baade 1951). However, a little known but remarkably prescient study employed newly invented PbS detectors to search for the Galactic center (Stebbins and Whitford 1947) using the 60-inch reflector 100-inch telescope at Mt. Wilson. They clearly detected the Milky Way bulge at 1.03 microns, and only just missed discovering the Galactic nucleus a full two decades before Becklin and Neugebauer (1968).




■ Fig. 6-5


Color-magnitude diagram of the Galactic bulge based on photographic photometry (Arp 1965). The spread in color genuinely reflects the range in metallicity and is not due to differential reddening over the relatively compact 1–4 arcmin of the Arp annulus. The same work clearly reveals the red giant branch of the metal-poor globular cluster NGC 6522, around which the annulus that included these stars was measured. Photographic plates in the R and I band obtained from Cerro Tololo resulted in a tight, well defined giant branch

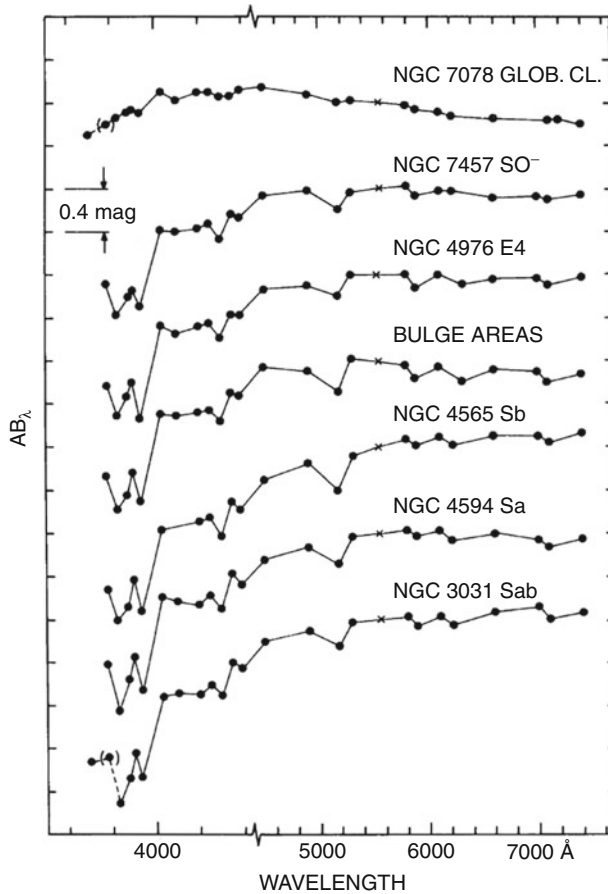
The discovery of large numbers of M giants in the Galactic bulge by V. M. Blanco (Nassau and Blanco 1958) had a significant impact at the Vatican Conference on stellar populations in 1957 – a turning point in the modern assessment of stellar populations. The careful consideration of the Galactic bulge at this workshop, based on the presence of both RR Lyrae variables (associated with old metal-poor stars) and M giants (associated with the disk), – led to the bulge being considered as a stellar population intermediate between the old disk and population II, a sophisticated assessment that can reasonably be said to stand to the present day. Unfortunately, the prevailing viewpoint of the time emphasized the presence of RR Lyrae stars (associated with metal poor globular clusters) in the bulge and consequently characterized the bulge as “old and metal poor.” van den Bergh and Herbst (1974) produced an interesting result based on photographic photometry from the Kitt Peak 4 m and Hale 5 m telescopes. This was the first work to reach the main-sequence turnoff in the Plaut window $(l, b) = 0^\circ, -8^\circ \approx 1,000$ pc from the nucleus. The van den Bergh and Herbst (1974) also found the earliest evidence of an abundance gradient in the bulge; and significantly, they found that metal-rich giants were more common at -4° (Baade’s Window) than at -8° . The author of this article realized that the actual photographic plates were themselves a valuable resource for proper-motion measurements; this initiated the proper-motion study of Vieira et al. (2007).

An important side note during this period is the controversy over supermetallicity. Spinrad and Taylor (1969) used early spectrophotometric observations to derive metallicities of field giants and stars in M67, finding abundances up to +0.6 dex higher than solar. The issue was controversial (See Rich 2008 for a more detailed history on the subject). Spinrad et al. (1969) even observed 3 M giants in Baade's Window at 30° elevation from Lick Observatory; Spinrad led numerous investigations into the spectroscopy of nearby galaxies. When supermetallicity was disputed, Peterson (1976) argued that the strong lines were due to CN blanketing affecting the stellar atmosphere; the subject was discredited. However, Branch et al. (1978) convincingly demonstrated that the field giant μ Leo was supersolar, using modern detectors and high-resolution spectroscopy. The early period of bulge investigations ended in the late 1970s.

The modern era of bulge research started with the operation of the Cerro Tololo Inter-American Observatory in the 1970s. Whitford (1978;  Fig. 6-6) showed that the integrated light of fields in Baade's Window has the spectral energy distributions of distant bulges and elliptical galaxies; this was a subtle effort that required a correction for the foreground disk light. Once it was possible to take photographic plates at the prime focus of the 4 m, Victor and Betty Blanco began a monumental survey of low-resolution spectra of late-type giants in the bulge and Magellanic Clouds, using a grating prism and IR-sensitive IV-N photographic plates. The most noteworthy results of this period were the grism surveys of late-type stars (Blanco et al. 1984) and the RR Lyrae surveys (Blanco et al. 1984; Blanco & Blanco 1997). In contrast to the Magellanic Clouds, which were found to be rich in carbon stars, the bulge lacked any carbon stars and was dominated by the M giants, with spectral types as late as M9. This population of luminous, long period variables, was recognized to be unique.

V. Blanco used the newly available red plates to construct a color-magnitude diagram in I , $R-I$, transforming the broad, undefined giant branch of the Arp (1965) photometry into a clearly defined sequence. This in turn made it possible to select a sample of bulge giants for spectroscopic study, pursued at Las Campanas using the just commissioned two-dimensional photon counter (Shectman 1984). The Las Campanas spectroscopy led to the study of 88 bulge giants that comprised the author's doctoral thesis (Whitford and Rich 1983; Rich 1988). At the same time, the first CCD imaging of the bulge employed the CTIO 4 m telescope (Terndrup et al. 1984; Terndrup 1988); this work revealed clearly the red clump and was the first to quantitatively demonstrate that the majority of bulge stars are old. Spaenhauer et al. (1992) availed themselves of the 30-year baseline between early 200-inch plates and plates obtained in the late 1980s, to obtain proper-motion measurements in the bulge.

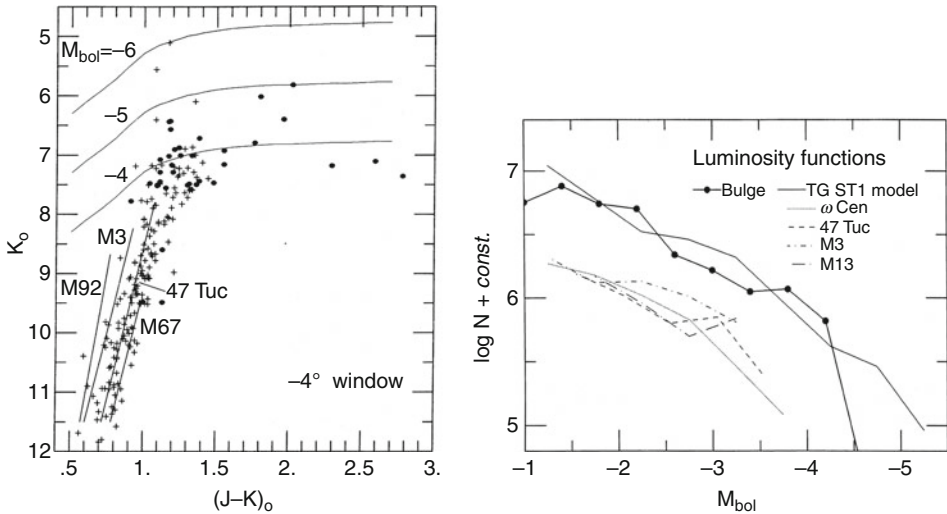
Blanco's sample of M giants enabled a major study of the bulge population in the infrared, led by Jay Frogel. Of the papers resulting from this work, Frogel and Whitford (1987;  Fig. 6-7) stands out as being of special significance, as it was the first infrared photometry of the bulge red giant branch (using the Blanco identifications and performed with a single-channel photometer). Subsequent studies of Blanco-selected M giants along the minor axis characterized gradients in photometric properties and line strengths (Terndrup et al. 1990; Frogel et al. 1990; Terndrup et al. 1991). In the 1990s, Sharples et al. (1990) reported the first multifiber spectroscopy of bulge K giants. McWilliam and Rich (1994) reported the first study of detailed abundances in bulge K giants. This work confirmed the high metallicities found by the early low-resolution spectroscopic studies and also showed that the bulge giants are enhanced in alpha elements that one would expect to be produced by core-collapse SNe (Wheeler et al. 1989).



■ Fig. 6-6

Spectral energy distributions of bulge fields near NGC 6522, globular clusters, and bulges in external galaxies (Whitford 1978). The bulge fields have a comparable spectral energy distribution (SED), and the abundance sensitive depression near the Mg I feature at 5170 (later used as a Lick index) is clearly comparable to the extragalactic bulges, along with the Ca HK break. An updated version of this plot is also given in [Fig. 6-51](#). The data were obtained by Whitford using the Harvard scanner at the CTIO 60-inch telescope

Following modern predictions of gravitational microlensing (Paczynski 1986), a number of projects were initiated to search for stellar gravitational lensing events. The MACHO survey began in the early 1990s and rapidly discovered microlensing events in the bulge (Alcock et al. 1997). The OGLE survey, using a 1 m telescope at Las Campanas Observatory, started shortly after MACHO and also found a high rate of microlensing events in the bar (Paczynski et al. 1994; Zhao et al. 1996). The EROS survey operates from a 1.5 m telescope at ESO while the MOA survey operates from New Zealand. The primary aim of bulge microlensing surveys shifted from dark matter studies to constraining the shape of the bulge and searches for events caused



■ Fig. 6-7

Color-magnitude diagram (*left*, variable stars solid points) and luminosity function (*right*) for the Galactic bulge, from Frogel and Whitford (1987). The lines of constant bolometric magnitude (*left*) apply to the bulge members only. The Galactic bulge red giant branch and luminosity function terminate at $M_{\text{bol}} = -4.5$, a characteristic of metal-rich old populations; note that the bulge contains stars more luminous than globular clusters owing to the short lifetimes of the most luminous phases (see also ● Fig. 6-14). The TG ST1 model refers to a “semitheoretical” model luminosity function of disk giants (Tinsley and Gunn 1976). The cool M giants observed in Frogel and Whitford (1987) were classified based on grating/prism spectra and images obtained at the CTIO 4 m telescope and had positions measured as well by Blanco et al. (1984). See Zoccali et al. (2003) for an updated version of this work, based on infrared imaging arrays

by lensing due to binary planetary companions. Microlensing surveys have also resulted in the discovery of rare, high-magnification events that can brighten bulge dwarfs sufficiently for high-S/N, high-resolution spectroscopy to be feasible. The latter effort continues to be important to the present (● Sect. 2.2).

Microlensing results continue to be one of the four pillars supporting the reality of the bar (gas dynamics, stellar kinematics, modeling of the projected light distribution, and microlensing). A current assessment by Hamadache et al. (2006) gives a compilation of microlensing results and models that support the existence of the bar, but this study argues that early microlensing work overestimated the optical depth.

While it is true that observed kinematics of HI and CO strongly support a bar-like potential, something that was understood even in the mid-1980s, optical observers took the bar seriously only after the asymmetry observed in the infrared images and measurement of the bar using red clump giants. About the same time as the microlensing results emerged, additional lines of evidence strongly supported the idea that morphology of the bulge is best described by a bar. Binney et al. (1991) from analysis of the gas dynamics, and independently, Blitz and Spergel (1991) from the infrared structure, called for a bar in the first quadrant. No special analysis was required to see that the bulge was vertically more extended in

the first quadrant; the early confirmation with star counts (Stanek et al. 1994, 1997) confirmed our present picture of a bar. The dramatic *COBE* images of the bar (● Fig. 6-1) provided convincing evidence that the bar is a reality in terms of the stellar distribution.

2 The Age and Population of the Galactic Bulge

The Galactic bulge is dominated by a globular cluster-age, metal-rich stellar population. The current debate is not over whether the bulge is mostly old but over what fraction, if any, may be in the range of ~few Gyr old. There are hints of such an age range from the evolved stellar content and, most recently, from the spectroscopy of microlensed dwarfs. Further, within 50 pc of the Galactic center, there is long-standing evidence for a very young stellar population; this is beyond the scope of this review, but we will discuss evidence for an old stellar population in the Galactic center.

The composition of the bulge offers evidence of rapid formation. Since McWilliam and Rich (1994), a number of studies find that the bulge is enhanced in alpha elements relative to the thin disk. This enhancement, traceable to early enrichment via core-collapse SNe (e.g., McWilliam 1997; Ballero et al. 2007), has been advanced as evidence that the bulge is old and enriched rapidly; this is corroborated by the evidence for pure r-process enrichment that we consider in ● Sect. 3.5. Although the alpha enhancement, reminiscent of the halo population, is circumstantial evidence for a starburst history, it lacks any absolute chronometer. Hence, inferences regarding the bulge age distribution are best drawn from the main-sequence turnoff and (less precisely) AGB stars.

Since the discovery of RR Lyrae stars in the bulge, there has been the certainty that a population of globular cluster-aged bulge stars exists. However, even at the stage of the Vatican meeting, there was ambiguity about the age distribution, it was recognized that the late spectral type M giants that dominated the bulge M giants were virtually nonexistent in globular clusters (Baade 1958). Baade's quote that opens this chapter indicates that spectroscopy and direct imaging of the M31 bulge was leading thinking in the direction of an old, metal-rich stellar population; this could not be verified at the time because the bulge of M31 was not resolvable with photographic plates (see ● Fig. 6-52).

The stellar content of the bulge may be regarded as similar to that of the populous globular cluster 47 Tucanae. While more metal poor than the mean bulge, 47 Tuc has a minority blue horizontal branch appended to a predominant red clump. In contrast with other populations, a substantial fraction of the bulge M giants are actually on the first ascent of the red giant branch because metal-rich stars are cool enough to exhibit TiO bands on first ascent. The second-ascent population is represented by the Mira variables, OH/IR and SiO maser stars. We will consider the second-ascent population later, as it may best represent the progeny of the trace intermediate-age population. The horizontal branch includes RR Lyrae stars but also includes a hotter HB population (Terndrup et al. 2004; see also ● Fig. 6-7 of Zoccali et al. 2003); some of these stars may be the counterparts of those stars responsible for the ultraviolet rising flux (UVX) in elliptical galaxies and bulges (O'Connell 1999). The most recently studied population in the bulge are the blue stragglers (Clarkson et al. 2011). While it is widely agreed that the bulk of the bulge population is as old as the oldest globular clusters, the mass fraction and age distribution of a possibly younger population has been a historically open question that remains actively debated to the present day.

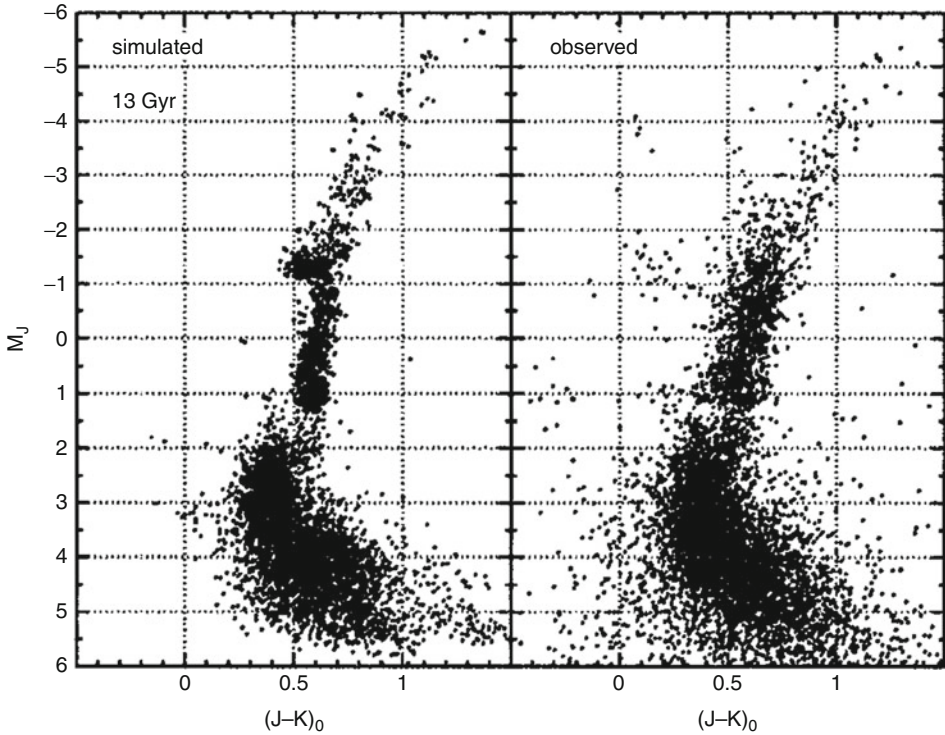
The stellar population toward a given field of the bulge consists of a foreground population of disk stars superimposed on the Galactic bulge population. Unfortunately, in most low-latitude fields, the foreground disk sequence overlies the old main-sequence turnoff. Terndrup (1988) reported the first CCD photometry for the Galactic bulge but did not correct for this foreground population. His work yielded the first solid detection of the red clump in the bulge, and overlaying isochrones made the case that the bulk of the stars at $b = -8^\circ$ and -10° are old.

Owing to the crowding, reddening, and superposition of disk stars, HST imaging remains the gold standard for turnoff-based age determination. Holtzman et al. (1993) reported the first Hubble Space Telescope (HST) photometry of the bulge, reaching several magnitudes fainter than ground-based work. However, a breakthrough was made by Ortolani et al. (1995), demonstrating that (1) the metal-rich bulge globular clusters NGC 6528 and 6553 have principal sequences (and therefore ages) almost identical to the well-studied globular cluster 47 Tuc and (2) when the red clump of the bulge is forced to align with that of 47 Tuc (eliminating distance and reddening differences), the sharp rise in the bulge field luminosity function matches a similar rise in NGC 6553, indicating room at most for $\sim 5\%$ of the population being younger than globular cluster age. The blue foreground population just brighter than the turnoff and observable over the whole of the bulge remained an issue of concern.

Feltzer and Gilmore (2000) showed conclusively that the blue foreground population arises in the disk, as the true bulge population shows a variation with Galactic latitude that is not mirrored by the disk. Two methods were then advanced to veto the foreground disk stellar population, and these remain as the standards to the present. Kuijken and Rich (2002) used proper motion to segregate the disk population: the distant bulge population appears to have $\mu_1 < -2.0 \text{ mas yr}^{-1}$, which is actually the apparent reflex motion caused by the solar circular velocity. Rejecting the disk stars leaves a globular cluster-like turnoff. The other approach, pioneered by Zoccali et al. (2003), subtracts a disk population at $(l, b) = 30^\circ, 0^\circ$ from the population observed at Baade's Window. Although the disk population is not identical due to possible abundance and age distribution gradients, this method overcomes the potential problem of kinematic bias. It also does not require the use of proper-motion measurements and hence may be applied to any bulge field of interest. The field subtraction method can be applied to a larger area of the sky and may prove especially useful for surveys employing the 4 m Dark Energy Camera (DECam) or LSST. The field subtraction method is shown in [▶ Fig. 6-8](#), and the proper-motion rejection method is illustrated in [▶ Figs. 6-9](#) and [▶ 6-10](#).

Neither approach can constrain the details of the age distribution of the bulge in the 5–12 Gyr range nor by itself give the age metallicity relationship. Strömgren photometry has the potential to do so and may be best employed by HST imaging. We can turn to either the bulge AGB population or composition, to give constraints on the intermediate-age population of the bulge. Relative age constraints might be affected if anomalous helium abundance affects part of the population, a possibility that is unlikely. Contiguous area integral field spectrographs like *MUSE* and *KCWI* (planned for VLT and Keck, respectively) have the potential to provide spectroscopic metallicities and radial velocities for stars in very crowded fields with HST imaging and proper motions; the ideal measurement of the age metallicity relationship and age distribution is within reach.

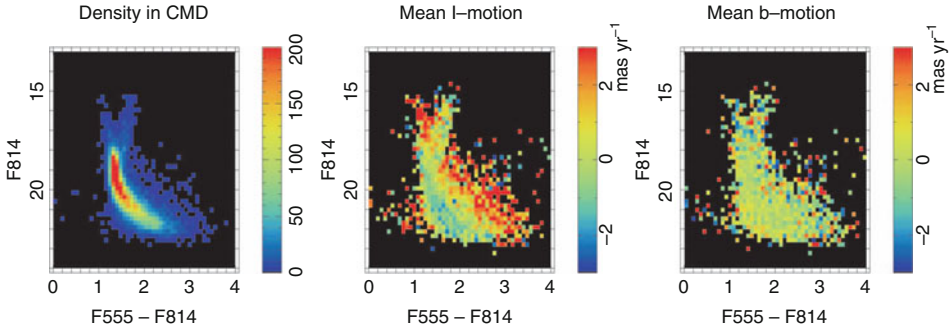
The SWEEPS field $(l, b) = 1.25^\circ, -2.65^\circ$ (Sahu et al. 2006) is one of the most intensively studied bulge fields (123-orbit HST integration); the original aim was to detect hot Jupiter transiting planet candidates for bulge dwarfs. In fact, some 16 transits were detected of which statistically $\sim 2/3$ are likely to be hot Jupiters (Sahu et al. 2006). Clarkson et al. (2008, 2011) employed this dataset and the proper-motion rejection method, to produce the strongest age



■ Fig. 6-8

Zoccali et al. (2003): Demonstration that the bulge is comparable in age to the halo globular clusters NGC 6528, by statistically subtracting the color-magnitude diagram of a field in the disk at $(l, b) = 30^\circ, 0^\circ$, which has little or no bulge population. The bulge population is observed with SOFI-LARGE on NTT, with the disk population statistically subtracted. NGC 6528 is originally observed with NICMOS but transformed to the ESO JHK system. The two CMDs are forced to align at the luminosity of the red clump to remove effect of different distance modulus and reddening at the red clump; The identical magnitudes of the main-sequence turnoff points demonstrate that any age difference between the field and cluster is very small (the cluster and bulge field have approximately solar metallicity). The same argument was used by Ortolani et al. (1995) employing optical colors and comparing the luminosity function of the bulge with that of the globular cluster

constraint for the bulge (► Fig. 6-9). Clarkson et al. (2011) show that many objects in the blue straggler region exhibit SX Phoenicis variable characteristics - placing even more stringent constraints on the fraction of a putative young population. Clarkson et al. (2011) find that only <3.4% of the bulge population can be younger than 5 Gyr. One weakness of the proper-motion separation method is that 6σ or greater precision is required for the proper-motion measurements; consequently the final acceptance sample for the bulge CMD (► Fig. 6-10) is $\sim 10\%$ of the original sample. However, when one recalls that two different methods of foreground disk correction (Zoccali et al. 2003; Feltzing and Gilmore 2000) give the same result, the case for an old bulge becomes especially compelling. The Clarkson et al. studies at present represent the what is currently definitive effort to constrain the age distribution of the bulge, although new datasets



■ Fig. 6-9

First demonstration of the proper-motion selection technique for constraining the age of the bulge (Kuijken and Rich 2002). The *left-hand panel* is the composite color-magnitude diagram (notice the turnoff point at roughly $F814 = 18$: two bright branches are the disk (*blue branch*) and the bulge subgiants (*red branch*)). The *right-hand panels* illustrate mean l and b proper motion. The bulge appears to stream opposite the solar motion, enabling one to reject foreground disk stars using the l proper motion (*middle panel*). The *light-green* sequence in the CMD has the turnoff and luminosity function characteristic of an old globular cluster; the stars in the *red sequence* have a CMD characteristic of the disk. Notice that the b proper motion does not yield any gain in terms of population separation (*right-hand panel*)

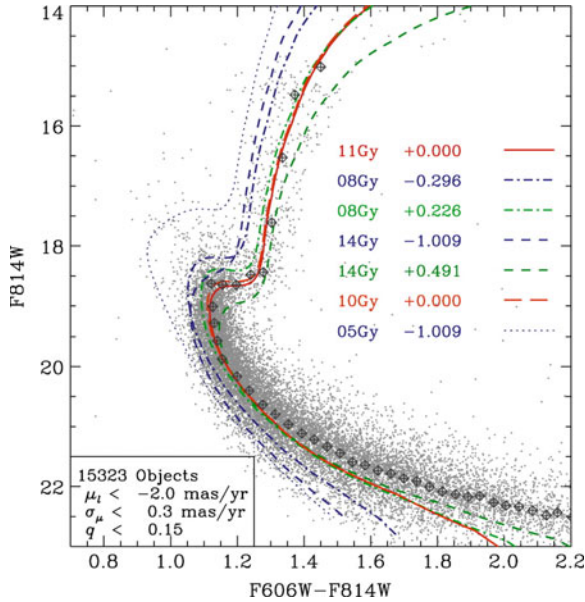
are expected to help examine biases still inherent even in the kinematic-separation methods (see Brown et al. 2010).

At the time of this writing, the class of globular clusters in the bulge with blue horizontal branches has been proposed to be possibly older than the mean age of the bulge and, conceivably, the oldest populations in the Milky Way (Ortolani et al. 2011). The relative age dating of these clusters and the bulge field stars will demand challenging observations for the faint stars near the turnoff (18–20 mag). Both improved distance constraints, and spectroscopic metallicity measurements will be needed to establish the age-metallicity relationship. The possibility that the bulge clusters and field might be the Milky Way’s oldest population is intriguing and worthy of further study.

2.1 Evidence for Minority Populations of Intermediate and Younger Age

As we have previously noted, the evidence supporting an old age for bulge stars outside of ~ 300 pc is very strong, but there has been a long-standing question about the luminous evolved stellar population and implications for a minority bulge component younger than 10 Gyr. Pioneering maps of the distribution of infrared luminous giants (Catchpole et al. 1990; [▶ Fig. 6-11](#) below) show that the most luminous giants are found in a flattened strongly concentrated distribution $< 1^\circ$ from the Galactic center.

Habing’s seminal investigation of the OH/IR stars as probes of the age and dynamics of the bulge appears to confirm this picture. Lindqvist et al. (1992a) found that the OH/IR stars

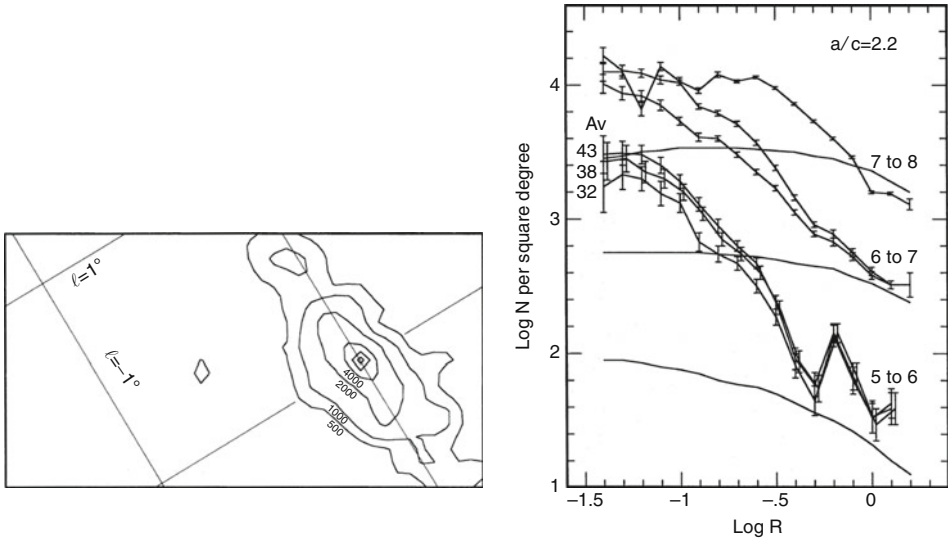


■ Fig. 6-10

Clarkson et al. 2008: Proper-motion-selected bulge objects from the SWEEPS field (l, b) = $1.25^\circ, -2.65^\circ$, with a tangent point 300 pc from the Galactic center; separation uses the mean proper-motion criteria similar to Kuijken and Rich (2002) but with a 6σ detection requirement imposed. The CMD was similar to the *lefthand panel* illustrated in Fig. 6-9, prior to cleaning. This CMD places very stringent constraints on the fraction of stars younger than 5 Gyr in the bulge. This CMD was divided into bins and the median computed (*diamonds*); below the MSTO, the uncertain binary fraction causes an artificial apparent age effect, so one focuses on the region above the MSTO for comparison. An α -enhanced, solar-metallicity isochrone at 11 Gyr represents the median sequence well above the turnoff. Also shown are sequences at metallicity $[Fe/H] = (-1.01, -0.23, +0.49)$ and ages (8, 10, and 14) Gyr to bracket the bulge population above the MSTO. Also shown is a very young, very metal-poor population (*dotted line*). A significant fraction, possibly all, of the apparently main-sequence stars brighter than the turnoff are blue stragglers (Clarkson et al. 2011). A red horizontal branch population can be seen at $F814W = 15.5$; the blue HB stars are too rare to appear in this CMD

with larger shell expansion velocities (candidate younger stars) are concentrated to the Galactic center; this finding persists to the present day (Sjouwerman et al. 2000; Fig. 6-12).

Turning to the Mira variable population, the picture is similar: Lloyd Evans (1976) and Whitelock et al. (1991) found Miras from $2^\circ < |b| < 8^\circ$ to exceed the 400-day period associated with younger ages (Feast 1963); this population is also considered in Whitelock (1992). Frogel and Whitelock (1998) also confirmed that the luminous long-period variables in Baade's Window have no counterparts in the metal-rich globular clusters. However, long-period variables are short lived and therefore rare, and their presence in the field does not imply that the field is younger than the clusters. A modern survey of Mira variables compiled from the OGLE II microlensing survey (Blommaert and Groenewegen 2007; Fig. 6-13) shows Miras

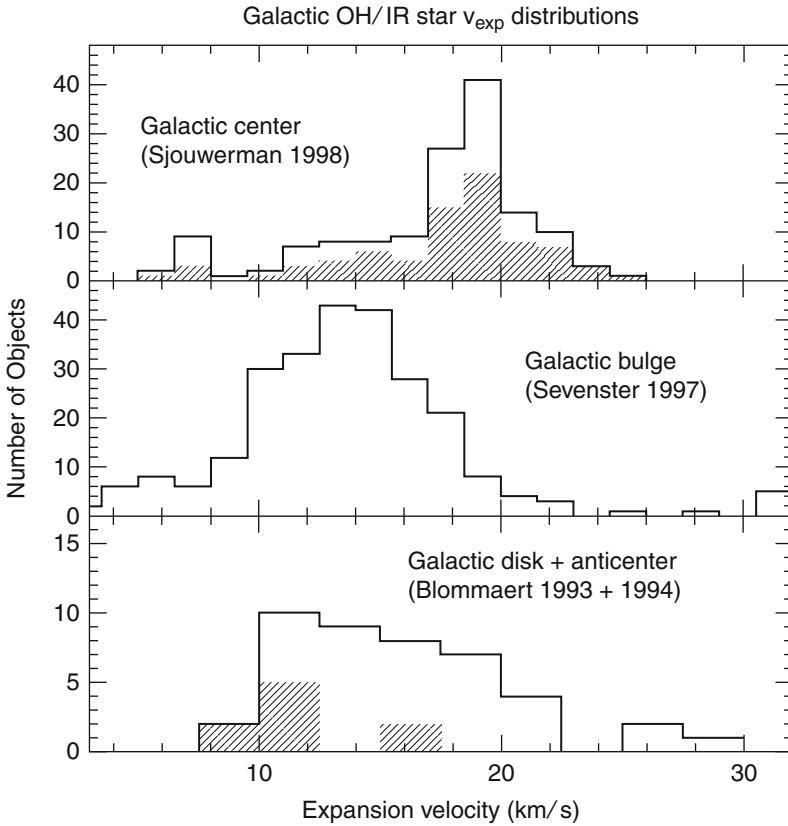


■ Fig. 6-11

(Left) Pioneering star counts with $6 < K < 5$ in numbers per square degree, from Catchpole et al. (1990). The observations were made with a single-channel detector. The flattened distribution of bright stars follows the nuclear “bar” of Launhardt et al. (2002) and is strongly concentrated to the nucleus on <50 pc scale. Contours of iso-star density are illustrated. (Right) Star counts for stars in various K-band luminosity ranges. The brightest (youngest evolved stars) are the most centrally concentrated; the fainter stars follow a shallower profile. X-axis in log degrees (Catchpole et al. 1990)

with period > 400 days are strongly confined to the inner 50 pc. This is in agreement with the picture from Catchpole et al. (1990) and from the OH/IR population. van Loon et al. (2003) analyze data from the ISOGAL project and find a trace population of young stars (as recent as 10^8 year) over the entirety of the bulge, even in Baade’s Window. In light of Clarkson et al. (2011) analysis of the blue straggler population, there is a reasonable concern that the main-sequence progenitors for the putative 100-Myr–1 Gyr-old evolved stellar population are not identified.

The debate over whether luminous AGB stars are young, especially at high metallicity, was crucial in the period following the Frogel and Whitford (1987) paper. Frogel and Elias (1988) showed that the bolometric luminosity of the brightest AGB stars in globular clusters increases with metallicity, and Guarnieri et al. (1997) found one LPV at $M_{\text{bol}} = -5$ in the solar-metallicity, old, bulge globular cluster NGC 6553 (► Fig. 6-13). The Galactic bulge has a stellar population $\sim 5 \times 10^9 L_{\odot}$, and therefore rare evolutionary phases are $\sim 10^4$ more common than in globular clusters. Luminous AGB stars evolved from old, metal-rich, stars are likely to be present. Further, the bulge hosts a proven (based on light curves) population of blue stragglers (Clarkson et al. 2011) that may evolve into luminous Miras (Greggio and Renzini 1990). It remains unclear how age and metallicity play off to produce a Mira population, and until theory becomes more precise, a definitive interpretation of the AGB content of the bulge, in terms of age, will not be feasible. ► Figure 6-14 suggests that the bulge hosts luminous, long-period Mira variables not

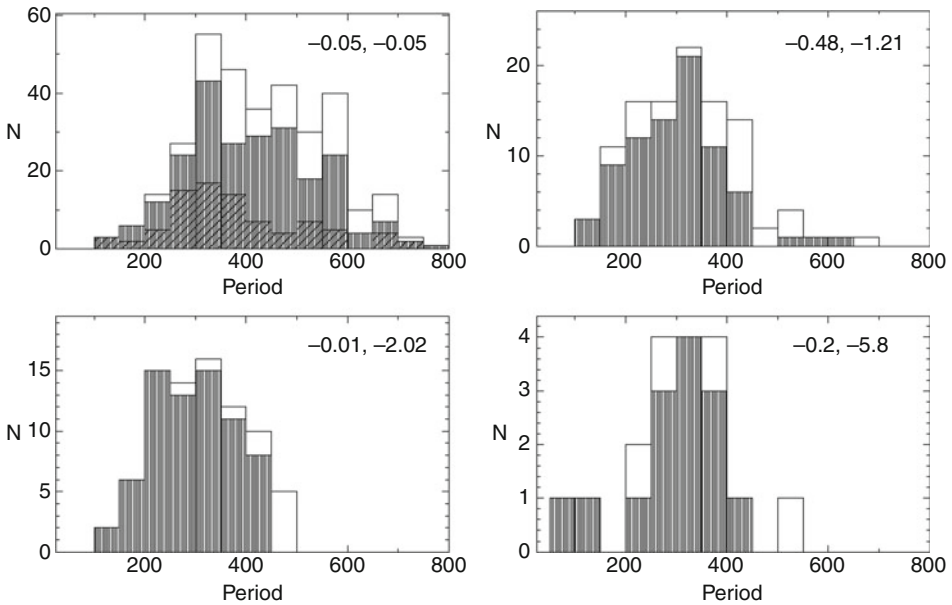


■ Fig. 6-12

Expansion velocity distribution for OH/IR stars (Sjouwerman et al. 1999). Higher expansion velocities may be related to higher mass or metallicity. The Galactic bulge noticeably lacks OH/IR stars with high expansion velocity (possibly more massive and therefore younger) found toward the Galactic center. This result is consistent with a concentration of more massive stars to the Galactic center

found in globular clusters, but since no globular cluster reaches the upper metallicity limit of the bulge, it is not possible to clearly define a period and bolometric luminosity limit beyond which an observed Mira variable is required to have massive progenitor.

A small number of AGB stars with the unstable element Tc have been found in the bulge; the presence of Tc requires a $1.5 M_{\odot} = 3$ Gyr progenitor (Uttenthaler et al. 2008), but such stars might evolve via the blue straggler channel. It is also noteworthy that the very small number of carbon stars found in the bulge are faint early R stars (Azzopardi et al. 1991; Tyson and Rich 1991), not the thermally pulsing AGB stars associated with intermediate-age populations. We conclude that based on the evolved-star evidence and the main-sequence turnoff data, *outside ~ 100 pc, the bulge is essentially a completely old stellar population with only a trace (at most a few percent) of intermediate-age stars.*



■ Fig. 6-13

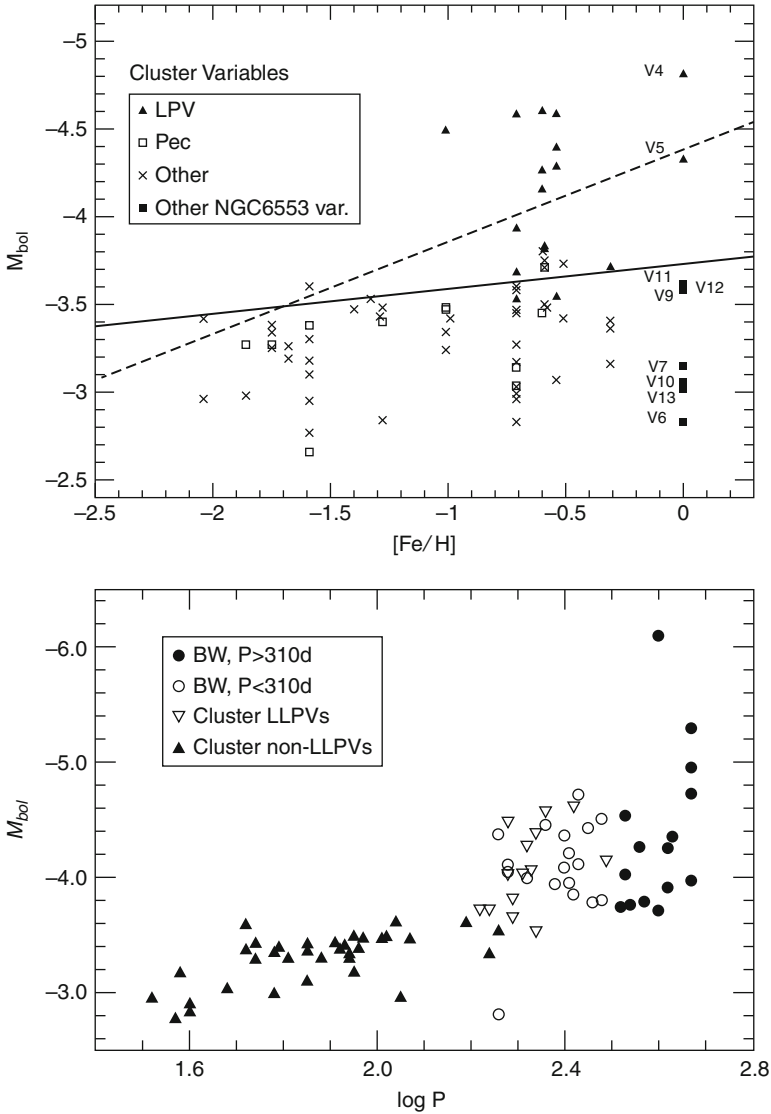
Period distribution of Mira variable stars from the OGLE II survey, from Blommaert and Groenewegen (2007). The filled histograms are for stars with $J - K < 2$; the (l, b) of the field is indicated in the upper right of each plot, progressing outward from the nucleus in Galactic latitude (recall Baade's Window is at ~ 500 pc and -4°). Notice that Miras with $P > 400$ days (likely to be young) are almost exclusively confined in the upper (innermost $< 1^\circ$) fields. The long-period Miras in the inner 100 pc likely arise from intermediate-age progenitors

The Galactic center region itself is special and has long been known to have extremely young stars (Morris and Serabyn 1996). The red clump is detected over the whole of the Galactic center, yet modeling of the luminosity function favors a continuous star formation history (Figer et al. 2004; ● Fig. 6-15). The presence of the red clump stars, however, strongly supports the existence of a population of stars older than 1 Gyr (and likely older) throughout the Galactic center. A combination of HST and ground-based AO imaging should soon give more detailed constraints on the star formation history of the Galactic center.

2.2 Microlensed Dwarfs: A Young, Metal-Rich Population?

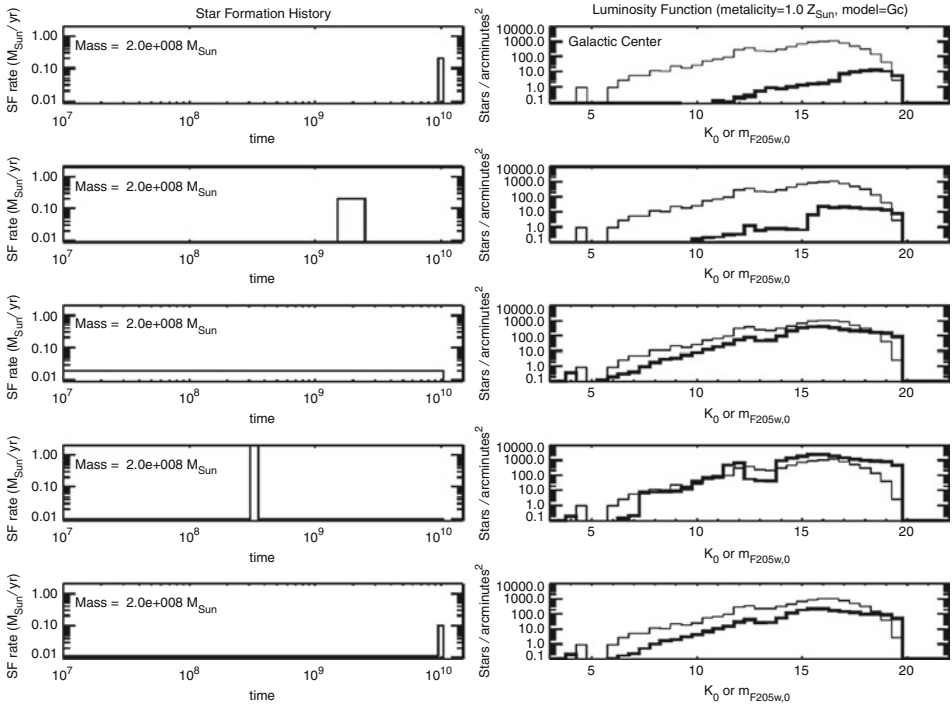
The bulge has proven to be an extraordinary laboratory for microlensing studies. Most of this work is beyond the scope of this review, as it uses the bulge stellar population to explore other problems such as lensing by planetary-mass companions. In the last few years, there has been great attention given to the targeting of highly amplified bulge dwarfs with high-resolution spectroscopy, first employed by Minniti et al. (1998). In some events, dwarfs can be boosted by more than 5 mag in apparent brightness so that dwarfs can be observed at high S/N.

Large microlensing surveys with automated alerts have made possible the scheduling of more target-of-opportunity spectra of highly amplified bulge dwarfs; at the time of this writing, some 40 such stars have been observed (Bensby et al. 2011). The technique is to use the spectrum



■ Fig. 6-14

The most luminous stars in the bulge are very long-period variables, but their presence does not require a minority young stellar population. (*Upper panel*) Luminous long-period variables are found only in the most metal-rich globular clusters, but the bulge reaches ~ 0.5 dex higher metallicity than the clusters. As Mira variables are short-lived ($\sim 4 \times 10^5$ year), the mass of the globular cluster population is only sufficient to produce the small numbers shown. Labeled variable stars are members of the solar-metallicity, old globular cluster NGC 6553 (figure from Guarnieri et al. 1997). *Solid line* is the RGB tip luminosity from Sweigart and Gross 1978; *dashed line* is the luminosity of models undergoing their first thermal pulse (approximate) from Renzini and Fusi Pecci 1988. (*Lower panel*) long period variables in metal-rich globular clusters and the bulge (Frogel and Whitelock 1998). The long-period variables in Baade's Window (bulge) also extend to higher luminosities; the most luminous variables may be evolved from the most metal-rich star binaries or may arise from a minority young population

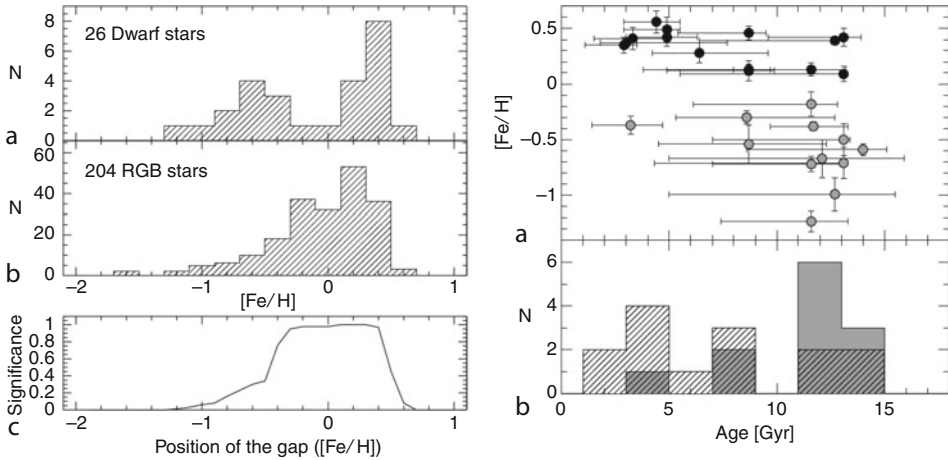


■ Fig. 6-15

(Figer et al. 2004); The Galactic center region is best fit with a continuous star formation history within the inner 50 pc. Model LFs (*heavy lines, right*) as a function of star formation scenario (*left*) for $Z = Z_{\odot}$ and the canonical mass-loss rates of the Geneva models, compared to the observed LF for the GC fields (*light lines, right*). The model counts have been multiplied by the completeness fraction of the observations. The models have been constrained to produce $2 \times 10^8 M_{\odot}$ over a circular area having $r < 30$ pc. Continuous star formation histories fit better than ancient bursts (Figer et al. 2004). The case for a young population in the Galactic center is secure based on other observations (see text)

only to give a “self-consistent” solution for abundance, effective temperature, and gravity. One can then place the star in the physical HR diagram and derive an age from the isochrones. Initial analyses of such cases yielded very young ages; at the metal-rich end, the method continues to derive surprisingly young ages.

Bensby et al. (2011), [▶ Fig. 6-16](#) below, found an apparently strong bimodality in their abundance distribution of 26 microlensed dwarfs; the bimodality is absent in an enlarged sample of 58 dwarfs (Bensby et al. 2012). Of greater concern is the apparent derived age spread at the metal-rich end, with a substantial population of metal-rich stars of age 2–6 Gyr; a result that persists in the larger 2012 sample. This population appears to be too large ($\sim 25\%$) to be consistent with the Clarkson et al. (2008, 2011) CMD-based age analyses nor is the result consistent with Zoccali et al. (2003). Bensby et al. (2012) consider the possibility of high He abundance as well as the population of bulge color-magnitude diagrams, and argue that the case for the ~ 20 – 25% of stars < 5 Gyr in age remains durable. Indeed, it is possible that some fraction of the most metal rich bulge stars are thin disk stars scattered away from the plane.



■ Fig. 6-16

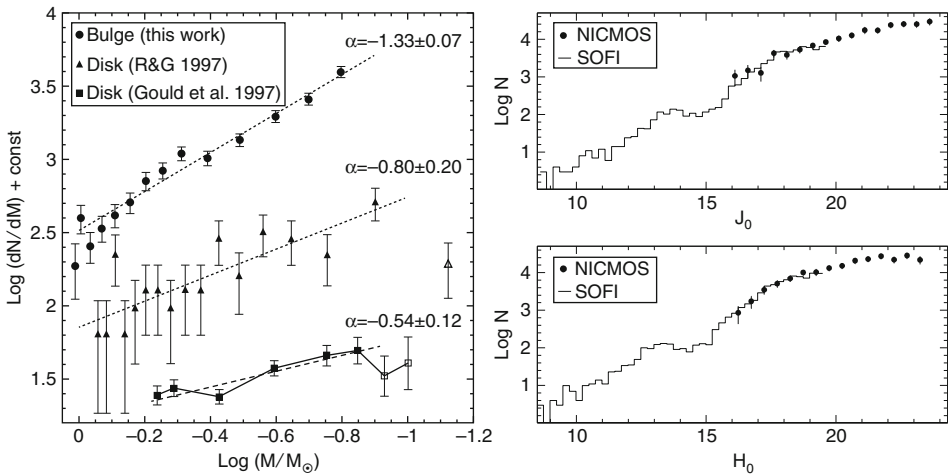
(Left) $[\text{Fe}/\text{H}]$ distribution of microlensed bulge dwarfs (Bensby et al. 2011). The lower figure shows a statistical test that suggests the two populations of dwarfs are real and not a statistical artifact, however a larger sample of 58 dwarfs no longer shows the bimodality (Bensby et al. 2012). (Right) Age distribution of microlensed bulge dwarfs, and age versus metallicity. These parameters are derived from the analysis of the very high-resolution, high-S/N spectra of these stars obtained during highly magnified microlensing events. The group of stars with age < 5 Gyr would have corresponding populations of turnoff stars that should be prominent in the color-magnitude diagram, and as well, there should be luminous AGB and C stars evolved from this population

If helium is enhanced in the bulge, photometric ages with standard isochrones would be incorrectly high, while spectroscopic ages would be measured as too young (Marín-Franch et al. 2010; Nataf and Gould 2012). It has recently been proposed by Nataf et al. (2012) that one may use the rare population of binary red giants to settle whether the candidate intermediate-age population is actually helium-enriched and old. The red giant masses at solar metallicity and greater vary from 1 to 1.4 M_{\odot} , with the greater masses corresponding to intermediate-age stars with normal helium abundance. It is clear that mass and metallicity measurements for binary populations in the bulge are of high priority.

➤ Section 8 briefly considers the formation history of the bulge. Additional constraints on the ages of stars are provided by the composition of the alpha and heavy elements. Broadly speaking, the abundance picture is more consistent with the bulge having experienced early, rapid formation (e.g., Ballero et al. 2007; Cescutti and Matteucci 2011), but the data do not rule out a small population of intermediate-age metal-rich stars.

2.3 The Luminosity Function

The present deep HST/NICMOS imaging of a field near NGC 6558 ($l, b = 0^{\circ}, -6^{\circ}$) has been extended to as low as 0.15 M_{\odot} corresponding to $H = 26$ (Zoccali et al. 2000; ➤ Fig. 6-17). The slope of $\alpha = -1.33$ is shallower than the Salpeter value of -2.35 but steeper than is found for the disk (Gould et al. 1997; Reid and Gizis 1997). Preliminary reductions of a (not much deeper) 100-orbit integration in the SWEEPS field at $b = -2.65^{\circ}$ find a similar result.



■ Fig. 6-17

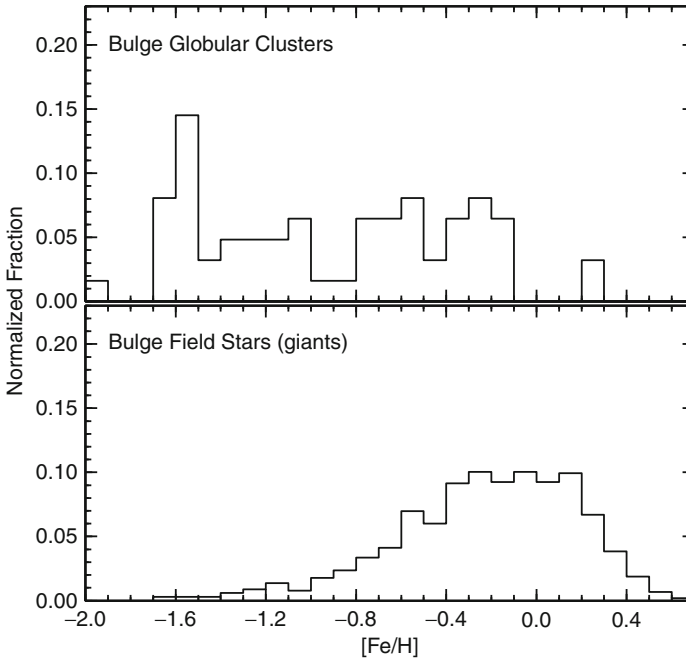
(Left) Faint bulge mass function based on HST/NICMOS imaging from Zoccali et al. (2000). One power law with slope $\alpha = -1.33 \pm 0.007$ fits the data over the full mass range; restricting the fit to $M > 0.5M_{\odot}$ gives a steeper slope of $\alpha = -2.0$. Two disk mass functions from Gould et al. (1997) and Reid and Gizis (1997) are more shallow. (Right) Complete empirical luminosity function in J and H bands, from Zoccali et al. (2003). There is a good agreement between the LF and an old stellar population model

It is interesting that the bulge appears to have a much more shallow IMF than the very steep low mass luminosity functions ($x = -3$) inferred for elliptical galaxies from the strength of the Wing-Ford FeH bands (van Dokkum and Conroy 2010). It will be therefore important to confirm the shallow IMF slope in other bulge fields with better data. It would be worthwhile obtaining actual spectra of bulge dwarfs and coadding them, to directly test the methods presently being used to constrain the mass function slope.

By integrating the IMF for the bulge from 0.15 to $100 M_{\odot}$ and adopting assumptions for binarity and stellar remnants (Zoccali et al. 2000), it is found that $L_{(VIJHK)} = (1.00, 1.13, 2.84, 3.67, 4.20) \times 10^5 L_{\odot}$ and $M/L_{(VIJHK)} = (3.67, 3.25, 1.28, 1.00, 0.87)$; the $M/L_K = 0.87$ is in good agreement with Kent (1992) value of 1 (Zoccali et al. 2003).

2.4 Globular Clusters

As in extragalactic systems, the inner galaxy globular clusters are associated with the bulge, not with the disk. Minniti (1995a) and Cotè 1999 find that the cluster system has the dynamics of the bulge. Although the clusters are somewhat consistent with the bulge dynamics, they extend to far lower metallicity than the stars (● Fig. 6-18). Further, if shattered clusters built the bulge field population, they did not leave the trace of Na-O anticorrelation or light element anomalies that characterize cluster stars. The GAIA satellite has the greatest opportunity to measure the connection between globular cluster and bulge field orbits. The clusters are alpha-enhanced and indistinguishable from their halo counterparts, save for the unique system Terzan 5 with its doubled horizontal branch, 0.5 dex spread in $[Fe/H]$, and distinct populations with differing $[\alpha/Fe]$ (Ferraro et al. 2009; Lanzoni et al. 2010; Origlia et al. 2011). Terzan 5 is a unique stellar



■ Fig. 6-18

[Fe/H] abundance distribution of bulge globular clusters (*upper panel*) and bulge giants on the minor axis (*lower panel*). The globular cluster population extends to *lower* metallicity than the bulge field and lacks the high-metallicity tail of the bulge field. The globular cluster population exhibits the usual alpha enhancement, with only Terzan 5 (Ferraro et al. 2009; Origlia et al. 2011) being unusual in hosting a very wide abundance spread of 0.5 dex in [Fe/H] (Cluster data courtesy of L. Origlia, figure by C.I. Johnson)

system and has been proposed as the tidally truncated remnant of a population of more massive galactic nuclei/primordial building block stellar systems.


3 Composition

This review will address abundance constraints from spectroscopy and will not consider helium, which must be inferred from a combination of star counts and stellar evolution models, either from the red clump (Renzini 1994; Minniti 1995b), the RGB bump (Nataf et al. 2011), or binary masses (Nataf et al. 2012). The complications of reddening and distance/membership uncertainty, along with uncertainty on the models, rarity of the populations, etc., make inference of the helium abundance a challenging measurement at present. Future missions like *GAIA* or *JASMINE* may yield parallaxes and proper motions for very large samples of bulge giants; comparison with those datasets may ultimately give helium abundance constraints for the bulge. Mass and composition measurements for binary red giants might offer another approach to constraining the helium abundance (Nataf et al. 2012). Diverse approaches will be required in order to get a constraint, and it is likely that the helium abundance will depend on [Fe/H].

A handful of Li-rich red giants have been detected in the bulge (the first being by McWilliam and Rich 1994). The origin of such extreme Li enhancement in $\sim 1\%$ of red giants is not explained and may reflect unusual evolution, deep mixing on the first ascent, or even ingestion of planets during the red giant phase. The bulge, with its dense population of red giants, is an ideal laboratory to survey for red giants with extreme Li abundance. Important details on the analysis techniques are also omitted from this review, but McWilliam and Rich (2004) and Fulbright et al. (2006, 2007) address the analysis and interpretation in greater depth than is possible here.

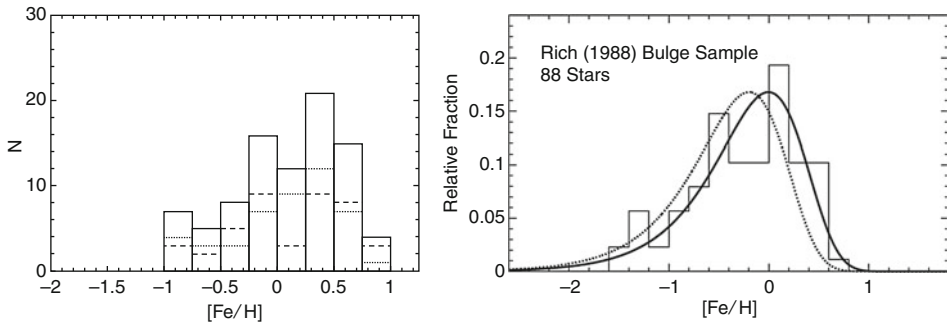
3.1 Optical Spectroscopy

The chemical composition of the bulge is critical to constraining its formation history and establishing its relationship to other stellar populations in the Galaxy. The light elements are believed to be predominantly produced in short-lived massive star SNe, the s-process in AGB stars, and the r-process in supernovae, likely core-collapse SNe (Wheeler et al. 1989; McWilliam 1997). The majority of iron is produced in Type I SNe, although 30–40% (depending on yields) is produced in massive star SNe. The composition gives constraints on the enrichment history of the bulge within its first Gyr and may also give insight into the initial mass function and masses of SNe, details that cannot be derived from other age constraints, even color-magnitude diagrams. Not all elements can be analyzed to yield this level of information. Mn, for example, shows evidence of a metallicity-dependent yield for both type I and II SNe (McWilliam et al. 2003). At the end of this section, the chemical evolution implications of these observations are considered. A deep theoretical interpretation of these results must await the completion of very large survey programs that are in progress.

As mentioned earlier, the definition of a red giant branch in R, I by Whitford and Blanco made possible the sample selection for the first low-resolution spectroscopic surveys (Whitford and Rich 1983; Rich 1988;  Fig. 6-19). The broad brush character of the bulge abundance distribution was laid out by this approach, which calibrated line strength as a function of against J–K color and $[\text{Fe}/\text{H}]$ for stars with abundances measured at high-resolution. The low-resolution spectroscopy gave a mean abundance of K giants in Baade’s Window of +0.2 dex, with the full range spanning -1.5 to $+0.5$ dex. Although with the benefit of high-dispersion, high-S/N spectroscopy, we find $\langle [\text{Fe}/\text{H}] \rangle = -0.1$ for the bulge Baade’s Window; the early work did capture the full range and character of the bulge abundance distribution.

Other large low-resolution spectroscopic studies followed (Minniti 1996a, b) but like all investigations were complicated by the ~ 15 – 20% contamination from foreground disk stars. Sadler et al. (1996) selected from the $(V-I)$ color-magnitude diagram and used low-resolution spectroscopy of a 400 K giant sample, confirming the general shape of the abundance distribution with 268 high-probability bulge members. The presence of stars of supersolar metallicity was expected, yet a source of some concern, until McWilliam and Rich (1994) used high-dispersion spectra to confirm the highest iron abundances and at the same time demonstrating that Mg and Ti are enhanced in the bulge, relative to the thin disk. McWilliam and Rich (1994) used spectra with $R \sim 17,000$ and $S/N \sim 50$ obtained using the CTIO 4 m telescope. CN is present throughout the optical spectrum of metal-rich giants, and only by using McWilliam’s synthesis of the complete optical CN spectrum was it possible to develop the list of iron lines unaffected by CN, which has made the subsequent studies possible.

Castro et al. (1996) analyzed the first spectrum of one of the most metal-rich bulge giants obtained with the HIRES spectrograph at Keck, showing it to have $[\text{Fe}/\text{H}] = +0.5$. Even with 8–10 m telescopes, high-resolution spectroscopy of bulge giants, especially in the more heavily



■ Fig. 6-19

Rich (1988) abundance distribution for the bulge, from low-resolution spectra calibrated using abundance standards. (Right) Corrected distribution from Fulbright et al. (2006) with Simple Model of chemical evolution (Searle and Sargent 1972) fits superimposed. Notice that the distributions have virtually no stars with $[\text{Fe}/\text{H}] < -1.5$ (this has been confirmed using in high-dispersion spectroscopy of $\sim 1,000$ bulge giants over a range of studies)

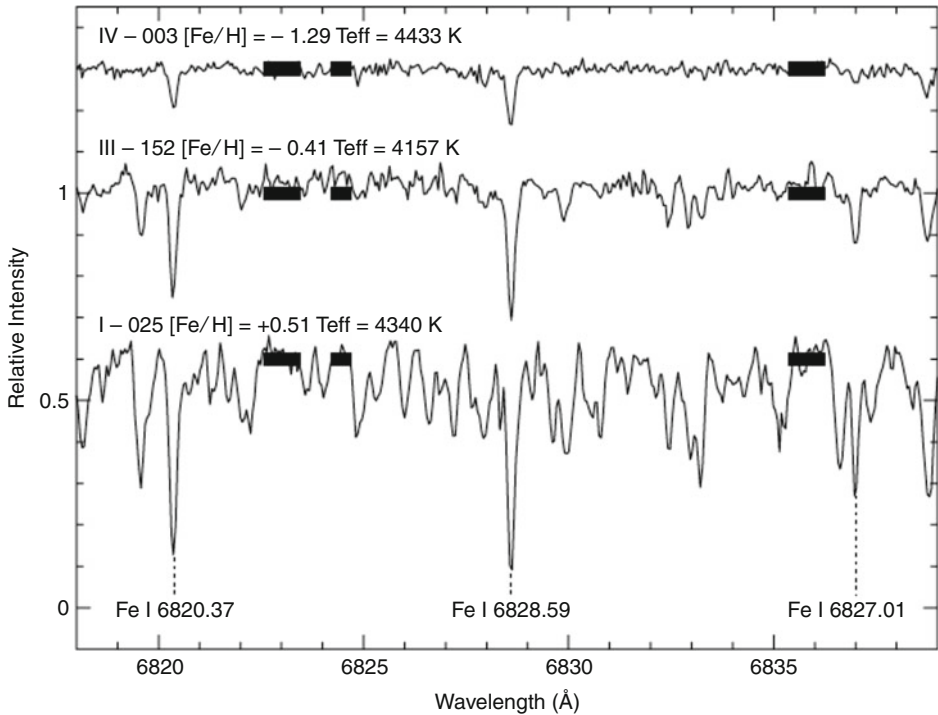
reddened low-latitude fields, is challenging. Fulbright et al. (2006) redetermined the fundamental stellar parameters of Arcturus and used Keck/HIRES spectra to derive a new abundance scale (● Fig. 6-20). This work gives a new line list of iron lines suitable for use in metal-rich bulge giants.

A consideration of all of the light elements by Fulbright et al. (2007) FMR07 confirmed that the alpha elements trend above the thin and thick disk distributions. They further divide the alpha elements into the hydrostatic alphas (O, Mg) which are produced in the burning layers of red giants and the explosive alpha elements (Ca, Si, Ti) that are produced in the core-collapse SN event. FMR07 finds that the explosive alphas define a narrow envelope in $[\alpha/\text{Fe}]$ versus $[\text{Fe}/\text{H}]$ with a scatter much smaller than that seen in the halo (● Fig. 6-23). Although the gas from which the bulge formed may have originated in the halo, the most metal-poor bulge stars exhibit far less scatter in $[\alpha/\text{Fe}]$ than the halo population in this study. It will be important to explore the bulge-halo connection in future work.

O and Mg follow different trends. $[\text{Mg}/\text{Fe}]$ is elevated at solar metallicity relative to the thin disk, a result found originally in McWilliam and Rich (1994). But $[\text{O}/\text{Fe}]$ is not as enhanced, leading to the proposal that the oxygen was not produced because the first generation of massive stars may have been rapidly rotating Wolf-Rayet stars and lost much of their envelopes in winds (Maeder 1992; see ● Fig. 6-23). McWilliam et al. (2008) and Cescutti et al. (2009) argued that there is also supporting evidence in the form of higher carbon abundances in the bulge; this is not supported by Ryde et al. (2009). Improved data may well resolve this issue.

The commissioning of the FLAMES multiobject spectrograph at VLT enabled a leap in multiplexing and S/N, with ~ 100 simultaneous spectra at $R = 20,000$ possible. An automated line-measuring code (DAOSPEC) was used to measure the iron abundance distribution of ~ 800 bulge K giants in fields at $b = -4^\circ, -6^\circ,$ and -12° , establishing the presence of an iron abundance gradient of $-0.6 \text{ dex kpc}^{-1}$ in the bulge fields outside of -4° (● Fig. 6-21). This study also confirmed that stars with $[\text{Fe}/\text{H}] < -1.5$ are extremely rare in the bulge.

Bulge fields of lower extinction, such as the Plaut field at $(l, b) = 0^\circ, -8^\circ$, have K giants that are just within the range of multiobject echelle spectrographs at 4 m telescopes. Johnson et al. (2011, 2012a, b) used the Hydra spectrographs at CTIO and WIYN to examine iron abundances



■ Fig. 6-20

Bulge giant spectra from FMR06, spanning a range in metallicity. The stars have similar physical parameters, other than the variation in $[Fe/H]$. The black bars are continuum points used for equivalent width measurement

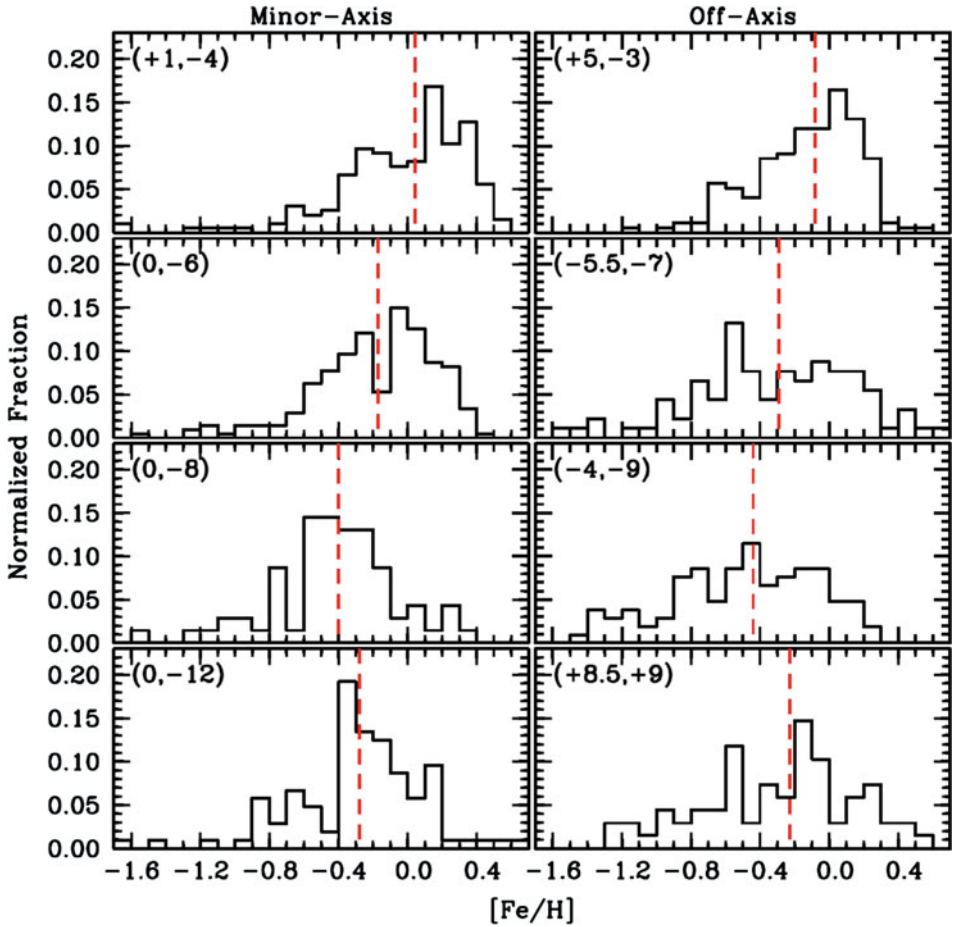
and elemental compositions of ~ 400 bulge K giants, up to mid-2012, with investigations being extended to off-axis bulge fields.

3.2 Infrared Spectroscopy

The advent of the cross-dispersed infrared echelle spectrograph NIRSPEC at Keck (McLean et al. 1998) enabled infrared spectroscopy of giants even in fields suffering from >10 mag of optical extinction; single-order high-resolution IR spectroscopy also became possible at VLT and Gemini.

Origlia and Rich have employed nirspec to carry out an extensive set of investigations in the bulge field and globular clusters using IR have employed nirspec to carry out extensive spectroscopy, establishing that composition can be derived from the infrared, with special success using the infrared OH lines near $1.6 \mu\text{m}$. Methods are given in Origlia et al. (2002). The bulge field and globular clusters are found to be similarly enhanced in $[\alpha/Fe]$, but the field has a narrow abundance distribution as in [Fig. 6-22](#) below (Rich et al. 2012).

IR spectroscopy of the bulge interior to Baade's Window finds no evidence of an iron abundance or $[\alpha/Fe]$ gradient with distance from the Galactic center, even within 50 pc (Rich et al. 2007b). This confirms the pioneering infrared low-resolution study based on Na and Ca lines

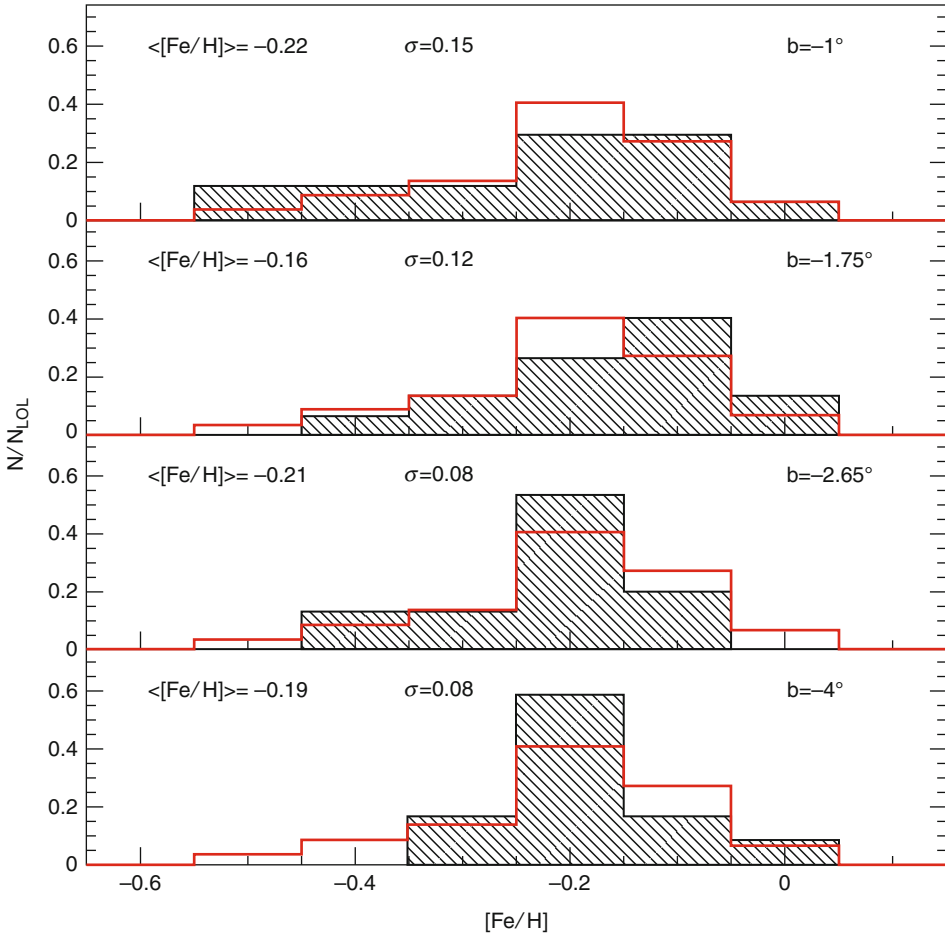


■ Fig. 6-21

The metallicity distribution from multiobject $R \sim 20,000$ spectroscopy for bulge fields is illustrated, with minor-axis fields on the *left* and off-axis fields on the *right*. Minor-axis fields at $(l, b) = +1, -4, 0, -6$, and $0, -12$ are from Zoccali et al. (2008), while that from $(0, -8)$ is from Johnson et al. 2011. The field at $(l, b) = +5, -3$ is from Gonzalez et al. (2011a), while the remaining off-axis fields are from Johnson et al. (2012b). A vertical metallicity gradient likely is present both on and off the minor axis, but no radial metallicity gradient is obvious

by Ramirez et al. (2000). While the alpha elements are elevated in all of the bulge fields, there is no evidence of a gradient in $[\alpha/\text{Fe}]$.

Ryde et al. (2009, 2010) use IR-derived abundances to argue that the bulge and thick disk follow similar abundance trends (see below). Fluorine at $2.2 \mu\text{m}$ may offer constraints on nucleosynthesis by massive AGB stars with ages in the 100 Myr range; resolutions $>40,000$ are required for this element. The extent of the study considers a small sample of bulge stars (Cunha et al. 2008), finding $[\text{F}/\text{O}]$ increasing with $[\text{O}/\text{H}]$ as is seen in the disk. Although F can be produced in both WR and massive AGB stars, the lack of strong s-process enhancement in the most metal-rich bulge giant with high F is not consistent with an AGB origin for the Fluorine. Future



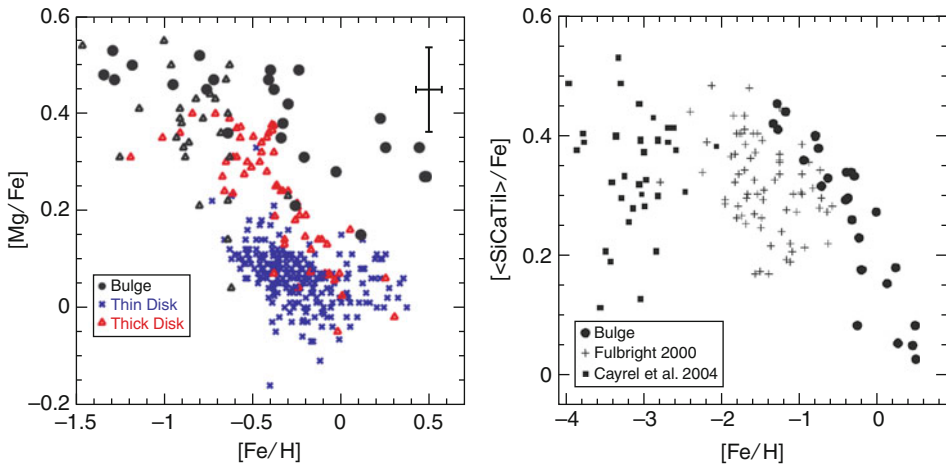
■ Fig. 6-22

Abundance gradient interior to Baade's Window (Rich et al. 2012) based on infrared spectroscopy of red giants; the study reaches within 140 pc of the Galactic center and shows no evidence for a gradient in $[\text{Fe}/\text{H}]$

F measurements may constrain formation scenarios that include wind-driven mass loss in WR stars and may be important as a test of the early massive WR population to explain the relative lack of enhancement of the oxygen.

3.3 Composition and Comparison with Other Populations

The early results of McWilliam and Rich (1994) showed an enhancement of alpha elements in bulge stars relative to the thin and thick disk populations, and these have been generally confirmed by subsequent studies (Fulbright et al. 2007; ● Fig. 6-23). Although the question of super metal-rich stars has been long debated (see Rich 2008), it has now proven beyond question that stars up to ~ 0.5 dex are found in the solar vicinity (Castro et al. 1997; Pompèia et al. 2003) and that NGC 6791 has $[\text{Fe}/\text{H}] \sim +0.4$ and are present in the bulge.



■ Fig. 6-23

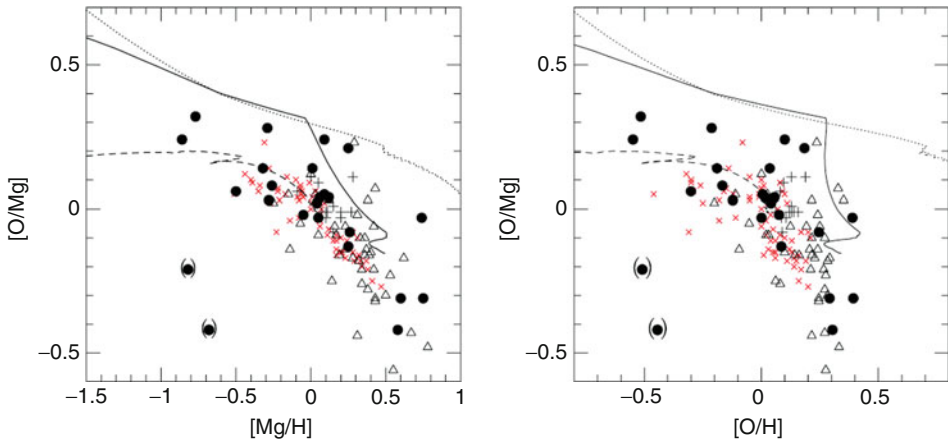
Mg and the “explosive alphas” from Fulbright et al. (2007). The bulge giants are clearly higher than the thin/thick disk in $[\text{Mg}/\text{Fe}]$; the tendency of $[\text{O}/\text{Fe}]$ to be less enhanced led to models (see Fig. 6-24) (McWilliam et al. 2008; Cescutti et al. 2009) in which mass loss in Wolf–Rayet stars early in the formation history of the bulge reduced the amount of oxygen produced. (Right) The bulge has little scatter in the explosive alphas (alpha elements produced in the supernova explosion as opposed to hydrostatic burning) $\langle \text{SiCaTi} \rangle$ at a given $[\text{Fe}/\text{H}]$, would appear to be consistent with the bulge having been significantly chemically homogenized than the halo—even at low (for the bulge) metallicity. The chemodynamical connection (if any) between the halo and bulge populations remains to be understood

Multiple studies (Johnson et al. 2011; Gonzalez et al. 2011a; Rich et al. 2012) find bulge stars to have $[\alpha/\text{Fe}]$ enhanced relative to the thin disk, over the entirety of the bulge, from 100 pc to the Galactic center, to 1 kpc distant (Fig. 6-25). It can be safely said that enhanced alpha abundances, especially at $[\text{Fe}/\text{H}] > 0$, are a signature of bulge membership.

Three issues remain contentious at the present time: first, whether the metal-poor bulge has a different pattern of elemental enhancement from the thick disk; second, whether a significant fraction of the metal-rich population is young (this latter point has been addressed in Sect. 2); and finally, whether the bulge has distinct chemodynamical subpopulations (e.g., Hill et al. 2011; Ness and Freeman 2012) based on structure, kinematics, etc.

Melendez et al. (2008) examined 19 bulge giants with high-resolution IR spectroscopy and comparable samples of thin and thick disk stars; their abundance trends from -1.5 to $+0.5$ dex are surprisingly similar, with the bulge stars showing higher $[\text{O}/\text{Fe}]$ only near solar metallicity. Ryde et al. (2009) also use infrared spectroscopy and find little distinction between bulge and thick disk in $[\text{O}/\text{Fe}]$. Alves-Brito et al. (2010) assert there is little difference between the bulge and thick disk; they used the published equivalent widths of the Fulbright et al. (2007) study and infrared spectroscopy of the same sample (Ryde et al. 2009) for CNO abundances. They added a new, high-resolution sample of thin and thick disk stars obtained with the same resolution and S/N as the Fulbright et al. bulge giants.

While the thin and thick disk stars appear similar from $-1 < [\text{Fe}/\text{H}] < 0$, at the metal-rich end, the Alves-Brito et al. (2010) sample has generally higher $[\alpha/\text{Fe}]$ than the thick disk and scatters to higher values. Further, the bulge extends beyond the metallicity of the thick disk, and



■ Fig. 6-24

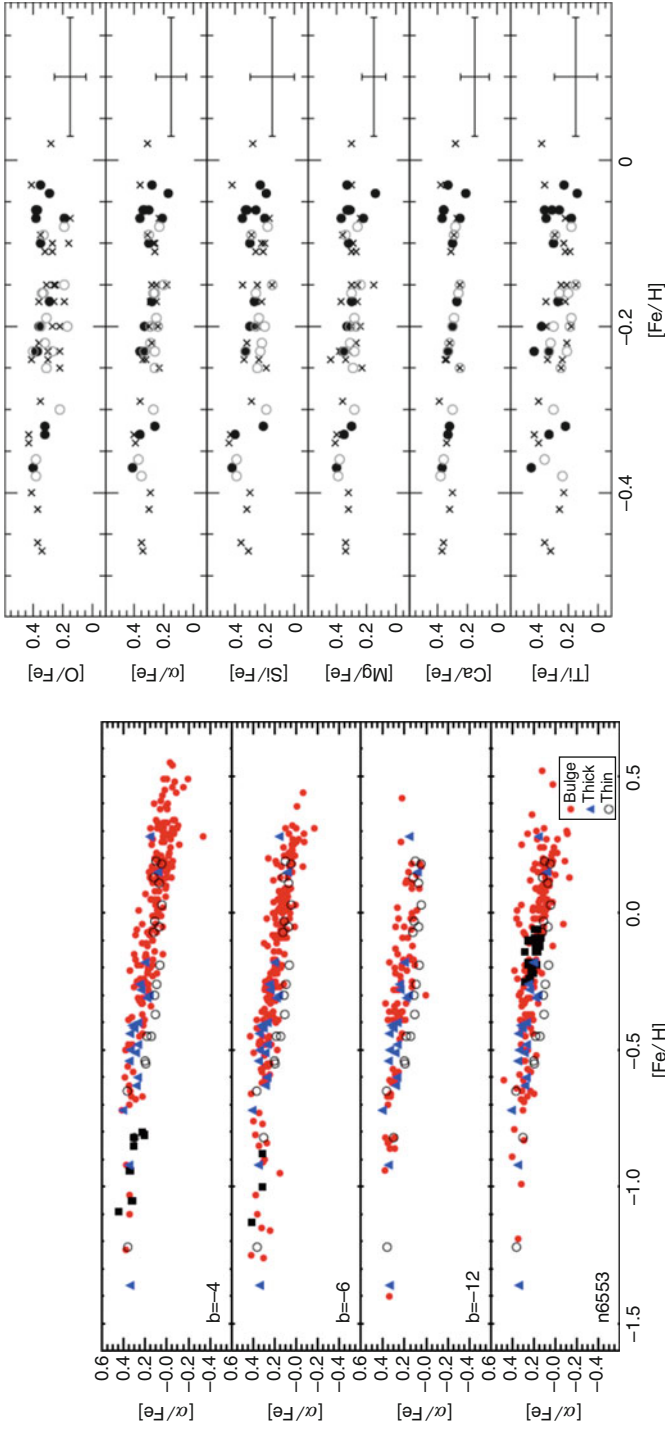
Oxygen is known to be less enhanced than Mg in the bulge. O/Fe versus [Fe/H] and [O/Mg] versus [Mg/H] from McWilliam et al. (2008). The *upper dotted curve* indicates results expected from nonrotating massive star nucleosynthesis in the early bulge (Ballero et al. 2007; yields, Woosley and Weaver 1995), while the *solid curve* employs yields from rotating massive ($M > 25 M_{\odot}$) stars (Maeder 1992) that shed mass before oxygen is produced from He and C in the hydrostatic burning layers. The *dashed curve* is a prediction for the solar vicinity using the Maeder models. Bulge data points are from Rich and Origlia (2005) (*crosses*), Zoccali et al. (2006) and Lecureur et al. (2007) (*open triangles*, low-S/N and high-S/N data), and Fulbright et al. (2007) (*filled circles*). Solar neighborhood data points were taken from Bensby et al. (2005) (*red crosses*). Additional discussions and details in McWilliam et al. (2008) and Cescutti et al. (2009)

has clearly higher [Mg/Fe] and [Ti/Fe], as was originally seen in MR94 (● Fig. 6-23); Bensby et al. (2011) find similar trends using microlensed bulge dwarfs. Johnson et al. (2011) also find that below [Fe/H] ~ -1 , the bulge bears some similarity to the metal-rich halo population. It is clear that based on $[\alpha/\text{Fe}]$, the bulge population is distinct from the thick disk at [Fe/H] > 0 . An issue in these studies arises in the small sample sizes and whether one can properly assign any non-bulge giants with [Fe/H] > 0 to the thick disk.

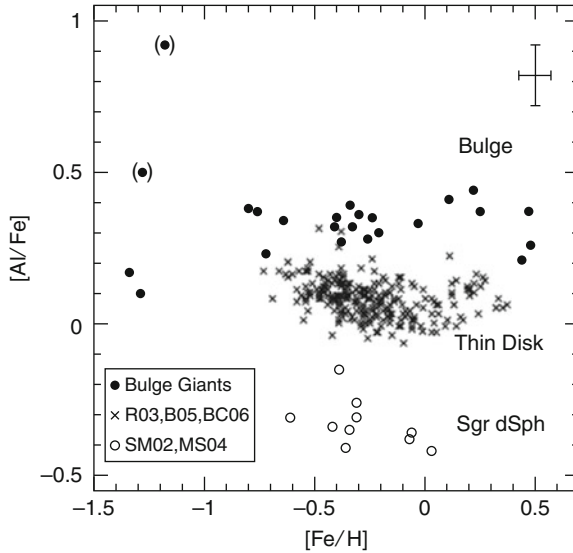
The alpha elements are not uniformly enhanced in the bulge and relative enhancements might potentially constrain the masses of primordial SNe responsible for the enrichment (cf. McWilliam 1998). ● Figure 6-24 shows a result from McWilliam et al. (2008) that models O and Mg trends might best be accounted for by an early generation of massive, rotating Wolf-Rayet stars. If the most massive stars in the proto-bulge shed their outer layers, there would be less oxygen (formed in the hydrostatic burning shells) relative to Mg, and increased carbon. Chiappini et al. (2011) propose a population of early, rapidly rotating massive stars, whose signature would be enhancements of Ba, Y, and Sr in some of the most metal poor bulge stars. More work is needed to test this intriguing model.

3.4 Na and Al

Some of the most significant differences between the bulge and other populations are found in the light odd elements Na and Al (● Fig. 6-26). Al differences are so significant that it may



■ Fig. 6-25 (Left) $[\alpha/\text{Fe}]$ versus $[\text{Fe}/\text{H}]$ for the outer bulge (Gonzalez et al. 2011a; optical spectra) and for the bulge $b < -4^\circ$ (right; Rich et al. 2012; based on infrared spectra). Enhancement of alpha elements relative to the thin disk is a fundamental characteristic of bulge giants and is observed over the whole bulge



■ Fig. 6-26

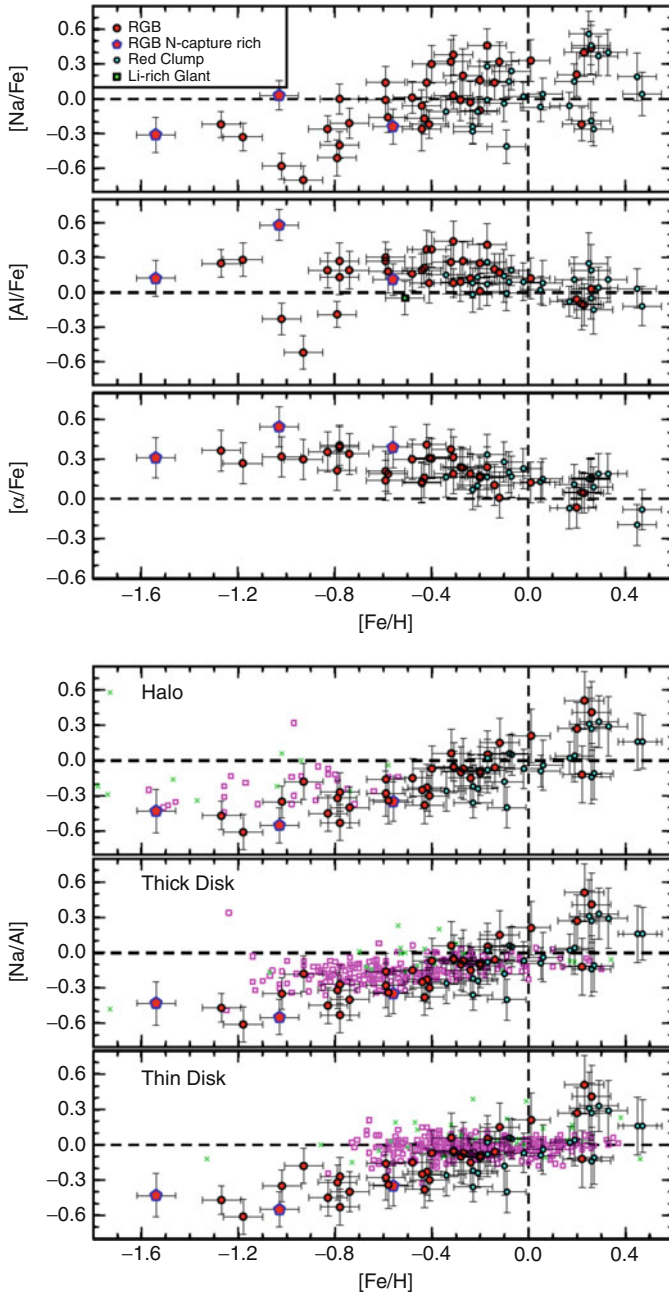
Strikingly different $[Al/Fe]$ versus $[Fe/H]$ trends in the Galactic bulge (*filled circles*), the Galactic thin disk (*crosses*), and the Sagittarius dwarf spheroidal galaxy (*open circles*) (Fulbright et al. 2007). Points in parentheses indicate giants identified as having O and Al abundances affected by envelope proton burning. The differences among stellar systems are great enough that Al may have value as a chemical tag useful in identifying stars in the bulge belonging to disrupted dwarf spheroidal systems

be usable as a powerful chemical tag to identify members of now disrupted globular clusters and dwarf galaxies. While Terzan 5 shows no Al-O anticorrelation (Origlia et al. 2011) it is crucial to constrain the fraction of field giants showing the Na-O anticorrelation and thus being members of shattered globular clusters; at present, we are lacking enough giants with Na and O measurements.

Fulbright et al. (2007) found a generally increasing trend of $[Na/Fe]$ with metallicity, while $[Al/Fe]$ follows an alpha-like trend. Other studies have found extremely high values of $[Na/Fe]$ in the bulge (Lecureur et al. 2007) but are likely affected by the difficulty of measurement for spectra of low S/N. Johnson et al. (2012a), ● Fig. 6-27 below, clearly show these trends and also the lack of a very strong $[Na/Fe]$ versus $[Al/Fe]$ correlation as is seen in some globular clusters. For both Na and Al, the bulge samples show trends that are not found in the thick disk (see Johnson et al. 2012a) as well. Al seems to be strongly dependent on environment (● Fig. 6-26).

3.5 Heavy Elements

The heavy elements are of interest because they potentially offer an independent set of age constraints. The r-process elements are very likely produced in massive star SNe, while the s-process elements are produced in AGB stars; there is the possibility of constraining the initial masses of AGB stars using specific elements, like Rb and Zr (e.g., Tomkin and Lambert 1983). Relatively few bulge stars have r- and s-process composition measurements. McWilliam and Rich (1994) reported the first Eu and s-process measurements, but the small sample size simply showed



■ Fig. 6-27

Trends of $[\text{Na}/\text{Fe}]$, $[\text{Al}/\text{Fe}]$, $[\alpha/\text{Fe}]$ versus $[\text{Fe}/\text{H}]$ for the Galactic bulge at $b = -8^\circ$ (Johnson et al. 2012a). The upper panel of the figure concerns trends of composition of bulge giants. The lower panel of figures compares the bulge with various other Galactic populations, illustrated using purple symbols. The strong $[\text{Na}/\text{Al}]$ trend is unique to the bulge population; notice the clear deviation from the thick disk trend. The low $[\text{Na}/\text{Fe}]$ for $[\text{Fe}/\text{H}] < -0.5$ is not observed in the thick disk

that the bulge was not atypical. McWilliam et al. (2010) published the first samples for Baade's Window giants, followed by Johnson et al. (2012a) (► Fig. 6-28). Eu/Fe has an alpha-like trend, although the origin of Eu is predominantly the r-process, and this trend should not be taken as demonstrating that Eu originates in the same SNe as the alpha elements. More likely, the trend reflects the metallicity-dependent yield of Eu. It is interesting that the bulge trend of [La/Fe] vs [Fe/H] is distinct from that of the thick disk, additional evidence that the bulge and thick disk have distinct chemical evolution histories.

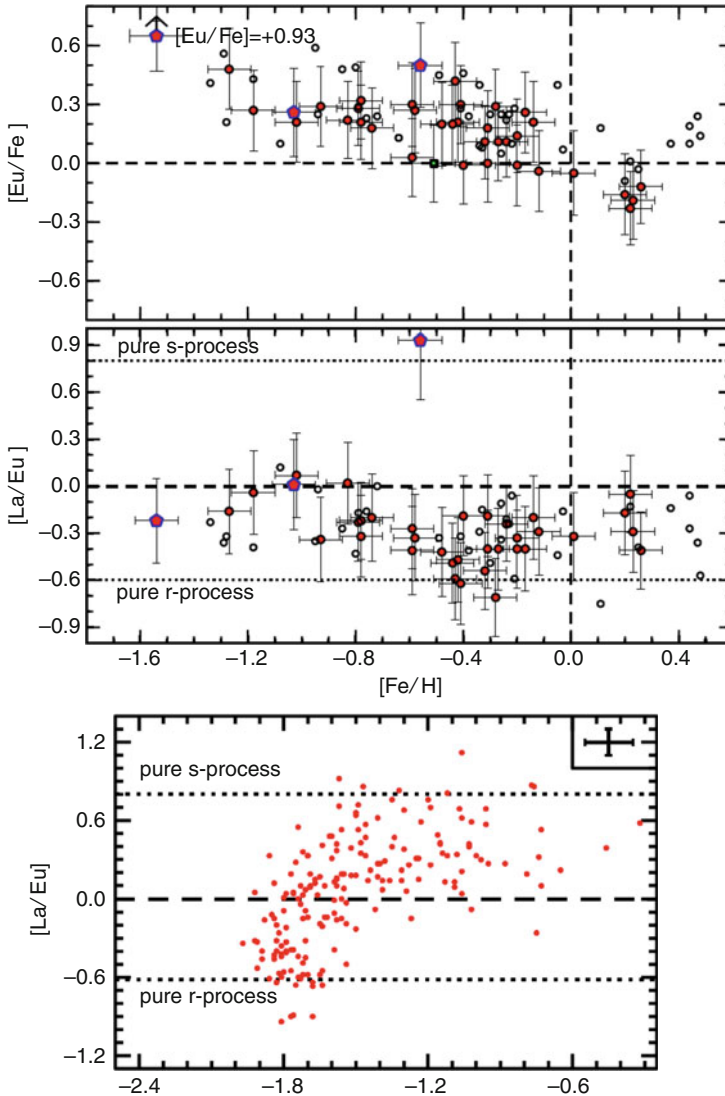
The trend of [La/Eu] vs [Fe/H] is a strong population diagnostic, as it is a proxy for the s/r ratio, and the bulge trend is distinctly lower than that of the thick disk (following a near pure r-process enrichment line) (► Fig. 6-28). There are also a handful of bulge giants with extremely high La abundances; these may be members of mass transfer binaries or dissolved globular clusters. The bulge does not show the strong increase in s-process exhibited by ω Cen (Johnson et al. 2012a; ► Fig. 6-28). The dip in [La/Eu] near [Fe/H] = -0.4 appears both in Baade's Window and the Plaut field and is therefore quite likely characteristic of the whole bulge. The relative weakness of the s-process at high metallicities is consistent with the rapid enrichment timescale implied by the trends in $[\alpha/\text{Fe}]$ versus [Fe/H]; the intermediate-mass AGB stars that are the source of the s-process-rich material (Gallino et al. 1998) appear not to have had the time to pollute the gas from which the most metal-rich bulge stars formed. If a population of young, metal-rich stars were present in the bulge, it would appear that there should be evidence of s-process contributions from AGB stars that would be expected on a 2–5 Gyr timescale. One need only examine the lower panel of ► Fig. 6-28 (ω Cen) to see appreciate how an extended enrichment timescale boosts the s-process in the younger population.

It is difficult to compare the bulge with the halo because so few bulge giants have [Fe/H] < -1.3, something we can confidently say because their fractional representation is so low in all of the bulge fields observed (e.g. Zoccali et al. 2008). However, [Na/Fe] and [Na/Al] are similar to values seen in the metal-rich halo (Johnson et al. 2012a). The most metal poor bulge stars do exhibit much lower scatter than halo stars, in the trend of the “explosive” alphas (Si, Ca, Ti) with Fe/H (► Fig. 6-23; Fulbright et al. 2007). If confirmed in larger samples, this would favor a different chemical evolution history that extends to the most metal poor members of the bulge. Much larger sample sizes will be needed to define how the bulge relates to the other populations on the basis of chemistry. While future surveys will likely discover stars with [Fe/H] < -3 in the bulge, it would be reasonable to expect these to actually be halo stars passing through the bulge, rather than representatives of the earliest bulge population.

Following consideration of kinematics and structure, the implications for the formation of the bulge are discussed.

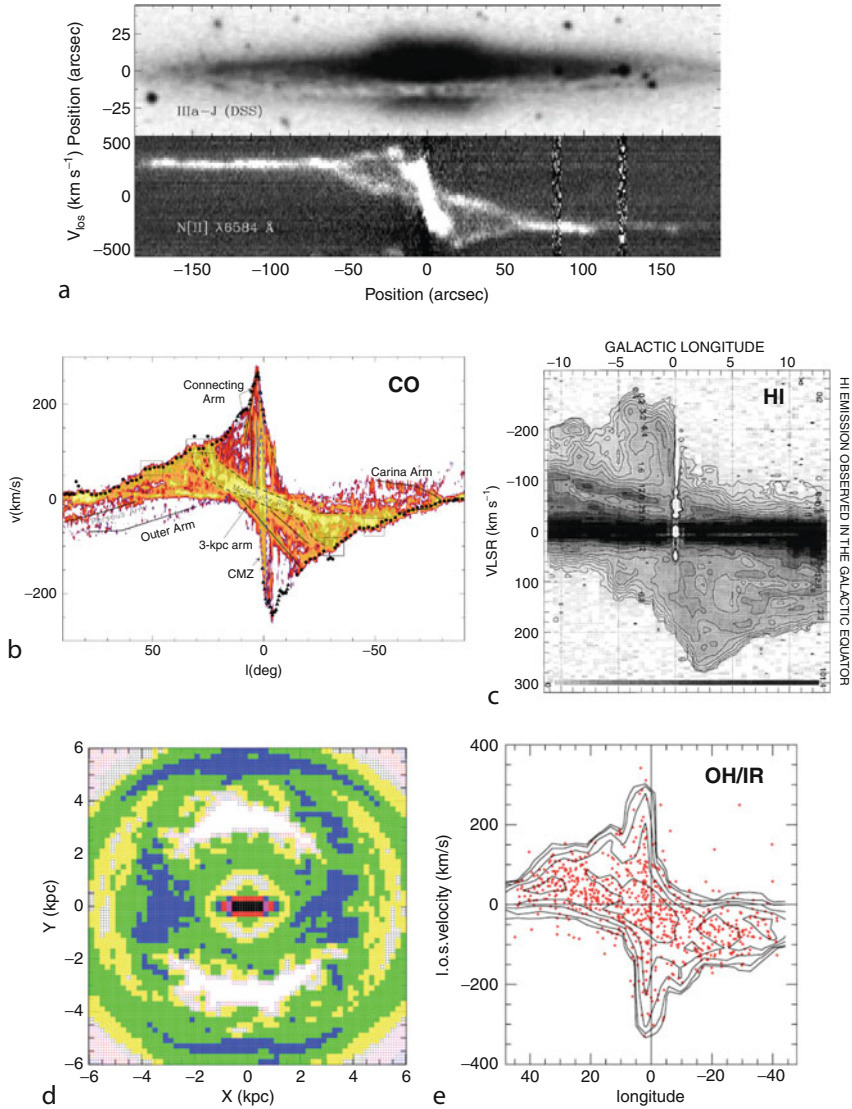
4 Kinematics

The dynamics of the bulge has been studied using both gas and stellar tracers. This review is concerned principally with the properties of the stellar bulge, but the kinematics of the gas has played an important role in early dynamical models (Kent 1992) and in early inferences requiring the presence of a bar potential (Burton and Liszt 1993). ► Figure 6-29 illustrates how one may “read” the complex distributions of gas in the longitude velocity plane and relate the gas to stellar radial velocities. It is possible to appreciate the disconnect between the radio observers who saw evidence of the bar and the optical observers that remained unconvinced until the COBE images were available. The hydrodynamic flow models are successful at reproducing a



■ Fig. 6-28

(Upper panels) Heavy element abundances in the bulge (Johnson et al. 2012a). These trends are consistent with a bulge formation timescale of < 1 Gyr and are distinct from those of the thick disk. Although Eu/Fe follows an alpha-like trend in the bulge, the decline is likely to be related to metallicity-dependent SN yields. Red symbols; Johnson et al. (2012a); open symbols; bulge giants in Baade's Window from McWilliam et al. (2010). The middle panel shows that the bulge follows an r-process enrichment pattern, consistent with enrichment so rapid that there was not enough time for AGB stars to contribute s-process elements, in contrast with the case for the thin disk and ω Cen (bottom panel from Johnson and Pilachowski (2010); notice that $[Fe/H]$ is lower than in the bulge). Notice how the extended formation time frame of ω Cen results in a dramatic change from r- to the s-process dominating heavy element production; presumably pollution from AGB stars is responsible. The bulge appears more to be a population in which the formation timescale was too rapid for the AGB stars to affect the chemical evolution. The pure r-process line is from Kappeler et al. (1989) while the pure s-process line is from Bisterzo et al. (2010)



■ Fig. 6-29

(a) Spectroscopy of the boxy-bulge galaxy NGC 5746 gives a visual guide that assists with the interpretation of the complicated HI and CO observations toward the Galactic bulge (Bureau and Freeman 1999). The geometry of how NGC 5746 and its ionized gas is viewed is comparable to our viewing perspective of the bulge perspective in the Milky Way. The structure in the ionized gas in (a) is caused by gas motions within the bar (Kuijken and Merrifield 1995). By way of comparison, it is possible to visualize the complicated structure of CO and HI in the Milky Way. (b) Annotated figure showing the longitude velocity diagram of the CO (1-0) emission (Dame et al. 2001) from Rodriguez-Fernandez and Combes (2008). (c) HI from Burton and Liszt (1993). Refer to (b) for CO and correspondence with various structures. (d) Model of the galaxy based on orbit modeling of OH/IR star velocities plotted; the *black bar* at the *center* shows the highest derived density of OH/IR stars (*red points*) in (e), overlaying the best fit model (Habing et al. 2006). Notice the correspondence of the OH/IR stars (*red points*) with gas kinematics contours in (e), see also SiO masers in Fig. 11 of Messineo (2002)

number of features observed in the gas (see [Fig. 6-30](#) below), and the use of a bar potential explained mysterious features in the gas l - v diagram, like the “expanding 3 kpc arm” that were of concern in the 1960s. Binney et al. (1991) employed the gas data to derive bar parameters similar to those favored today, at a time when the discovery of the bar from the early infrared data (Blitz and Spergel 1991) was unexpected. Englmaier and Gerhard (1999) also showed that the terminal velocity implied by the COBE bar agrees with that predicted from gas flow models. Modeling of the OH/IR stars also gives a pattern speed for the bar $\Omega = 59 \pm 5 \pm 10$ (sys) $\text{km s}^{-1} \text{kpc}^{-1}$ for $R_0 = 8$ kpc and $V_0 = 220 \text{ km s}^{-1}$ (Debattista et al. 2002); Gerhard (2011) review recent work on the bar pattern speed, including agreement based on constraints from stellar streams in the solar vicinity (Dehnen 2000).

4.1 Stellar Radial Velocity Surveys

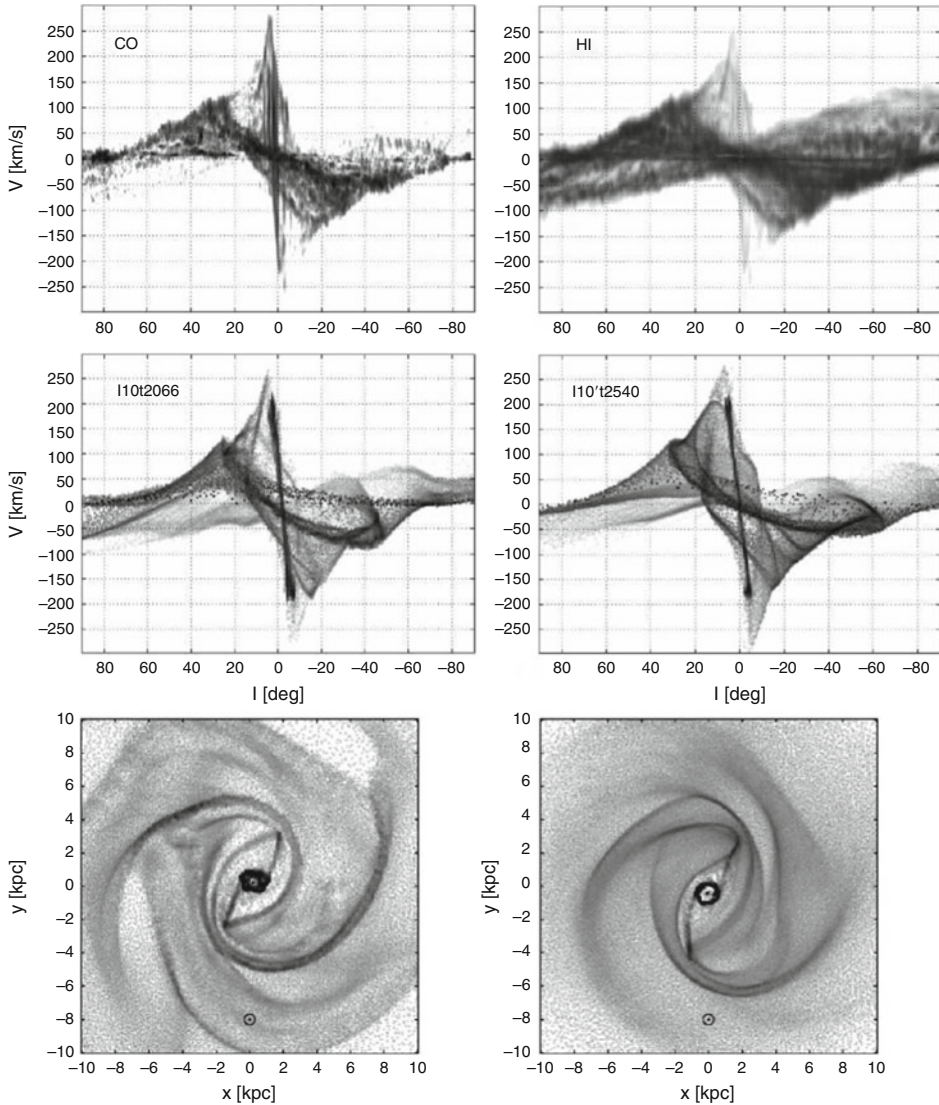
Early kinematic surveys of the bulge K and M giants measured a velocity dispersion of 110 km/s in Baade’s Window (Mould 1983; Rich 1990). These results were confirmed by Sharples et al. (1990) who undertook the first multiobject spectroscopy in the bulge, an important milestone in the study of kinematics. These investigations enabled the oblate rotator bulge model of Kent (1992), which included the first infrared results from the Spacelab mission (Maihara et al. 1978). An updated consideration of bulge minor-axis kinematics (prior to the BRAVA survey) from the nucleus to the edge of the bar is given in [Fig. 6-31](#) below (Tremaine et al. 2002). With the demonstration of the bar (Blitz and Spergel 1991), it became necessary to contemplate a rapidly rotating bar model for the bulge. Zhao (1996) published his rapidly rotating bar using the Schwarzschild method to populate orbit families in a self-consistent manner.

A significant breakthrough in the measurement of kinematics off the minor axis came with the K giant studies of Minniti et al. (1992) that yielded the first bulge rotation curve and hundreds of OH/IR stars (Sevenster et al. 1997) and planetary nebulae (Beaulieu et al. 2000). The PNe kinematics in Beaulieu et al. (2000) was fit by bar models, including that of Fux (1999) illustrated in [Fig. 6-28](#). While PNe are relatively rare, the Beaulieu et al. (2000) study was the largest survey of bulge stellar kinematics at that time, as the PNe covered a wide range in l and b (one may note the good agreement with the stellar rotation curve in [Fig. 6-33](#)).

The OH/IR stars are observed largely in the plane and include many sources near the Galactic center. Referring back to [Fig. 6-28](#), the l - v diagram of OH/IR stars has been modeled using orbit reconstruction (Habing et al. 2006) to demonstrate the likely presence of a bar, as well as to illustrate the presence of the corotation resonance at 3.3 kpc. Note that the SiO masers exhibit a similar radial velocity distribution that overlays nearly perfectly Damet et al.’s 2001 CO l - v distribution (Messineo 2002).

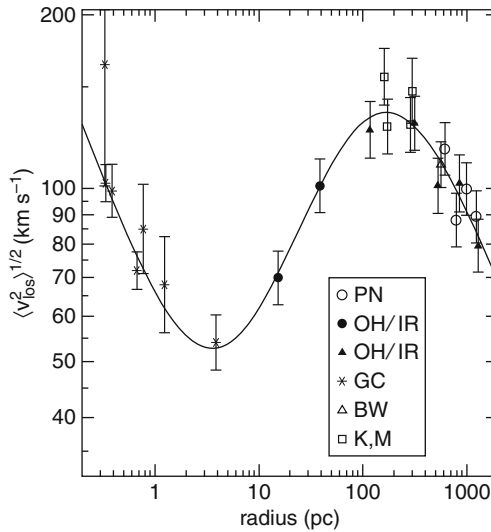
The advent of the Two Micron All-Sky Survey (2MASS) provided a virtually unlimited sample of giants spanning the entire Galactic bulge, largely accessible regardless of field extinction. This enabled Rich et al. (2007a) to initiate the *Bulge Radial Velocity Assay* or BRAVA utilizing the 100-fiber hydra spectrograph at the 4 m telescope at the Cerro Tololo Observatory.

[Figure 6-32](#) shows the color-magnitude diagram of the complete sample of 9,500 BRAVA giants observed along with the targeted fields. Because BRAVA surveyed the bulge in a grid spanning both latitude and longitude, it has become possible to easily investigate the rotation field perpendicular to the plane using a finely sampled set of velocity probes. The BRAVA survey shows that the bulge departs from pure “solid body” rotation (Howard et al. 2008) and exhibits cylindrical rotation (Howard et al. 2009; Rich et al. 2008, 2009), a characteristic of pseudobulges (Kormendy and Kennicutt 2004). The rotation field for the -4° and -8° slices is identical.



■ Fig. 6-30

(Fux 1999): Confrontation of a selection of two N-body models of a galaxy with a COBE-like bar, with observed gas kinematics. Refer to ► Fig. 6-29 to visualize how the observations of CO and HI relate to the galaxy. Note that these good fits are selected from a range of models and reflect gas motions in the bar potential. *Top*: ^{12}CO and HI I - V diagrams integrated over $|b| < 2^\circ$ and $|b| < 1.25^\circ$, respectively; the data are from Dame (1999) for the CO, and Hartmann and Burton (1997), Burton and Liszt (1978) and Kerr et al. (1986) for the HI. *Middle*: synthetic I - V diagrams of models I10t2066 and I10t2540 for a bar inclination angle 25° , including all particles within $|b| < 2^\circ$. *Bottom*: face-on projections of the gas spatial distribution in these models, rescaled as such to put the observer at $(x, y) = (0, -8)$ kpc (⊙ symbol). In these units, corotation lies at $RL = 4.5$ kpc. The model on the *left* reproduces almost perfectly the connecting arm, while the model on the *right* provides a fair global qualitative agreement to the data. Although the gas dynamics require a bar potential, but demonstrating the bar shape with using star counts has been an important step



■ Fig. 6-31

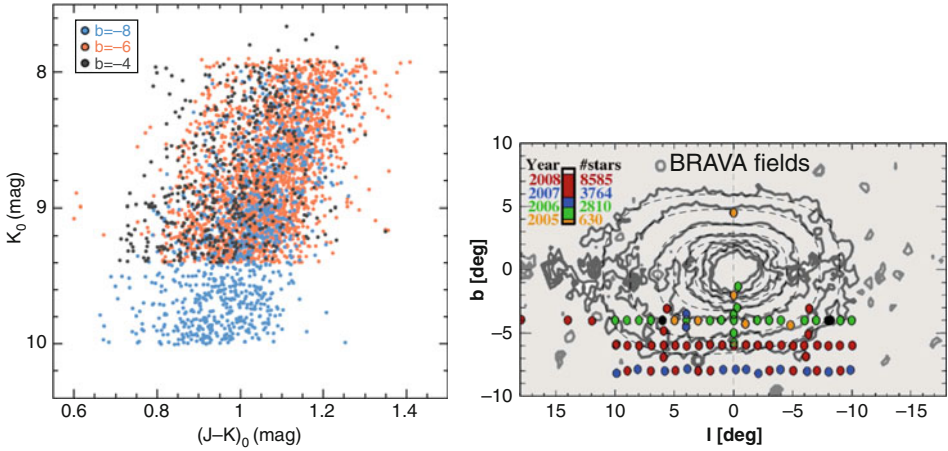
RMS line-of-sight velocity dispersion in the Galactic bulge, from the nucleus to the outer bulge (Tremaine et al. 2002). Galactic center stars (Genzel et al. 2000), PNe (Beaulieu et al. 1999); OH/IR stars (Lindqvist et al. 1992a; Sevenster et al. 1997); BW = giants in Baade’s Window (Terndrup et al. 1995); K, M = giant stars (Blum et al. 1994; Blum 1995). Filled symbols indicate observations closer to the Galactic plane; open symbols indicate observations away from the plane. The curve is a fitting function. The figure follows an earlier compilation of bulge velocity dispersions from Kent (1992)

BRAVA covered most of the southern half of the bulge, and none of the fields show clear evidence for cold streams. The most significant result indicates that the fraction of the bulge mass in a “classical” non-barred configuration must be <8% of the disk mass (Shen et al. 2010) (► Fig. 6-32). Analysis of new data by Kunder et al. (2012) confirms the cylindrical rotation for the -6° field as well.

Shen et al. (2010; ► Figs. 6-33–6-36) fits an evolving N-body disk model to the *BRAVA* dataset, finding that the radial velocity data are consistent with the bulge having formed from a bar that has undergone buckling. The model also leaves very little room for a “classical” slowly rotating metal-poor bulge, but it is noteworthy that theoretical models (Shen et al. 2010; Saha et al. 2012) predict that a rapidly rotating bar can spin up a classical bulge and would even be capable of forcing cylindrical rotation in a classical bulge.

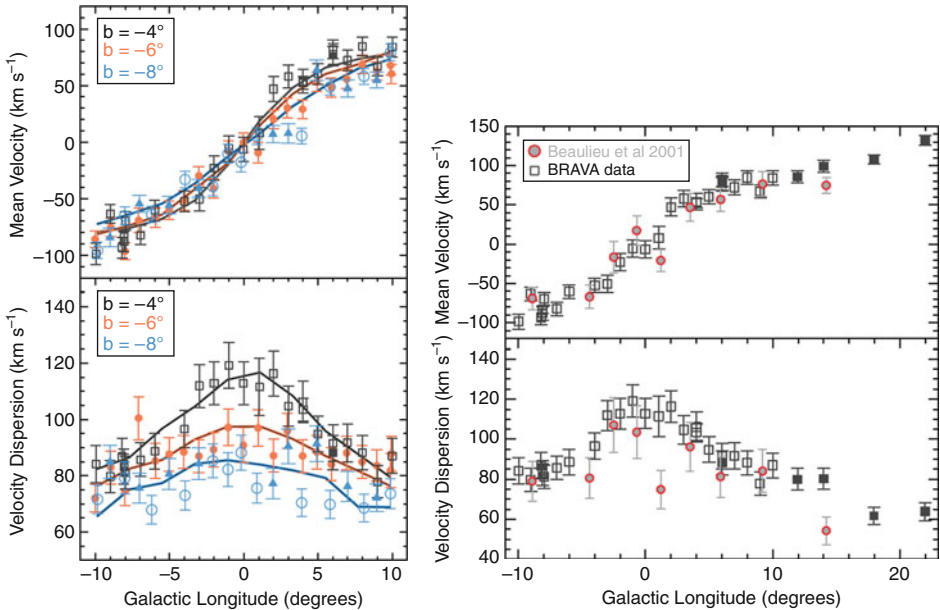
The new self-consistent bulge model – the update of the Zhao (1996) model (Wang et al. 2012) – fits the radial velocity rotation field well, but predicts larger than observed proper motions. The Wang et al. (2012) model has a pattern speed of 60 km/s/kpc, disk mass of $10^{11} M_\odot$, and bar angle of 20° for an adopted bar mass of $2 \times 10^{10} M_\odot$. The fraction of chaotic orbits ranges from 80–93%, depending on the integration time.

An additional result from the *BRAVA* study places the Milky Way in the Binney plot (► Fig. 6-52) with the bulge plotted near the well-known “peanut-shaped” bulge galaxy, NGC 4565, lying slightly above the oblate rotator model line. It is worthwhile recalling that the posited “X-shape” feature in the bulge (McWilliam and Zoccali 2010; Nataf et al. 2010; Saito et al. 2011) is in fact consistent with the rapidly rotating, N-body bar models of pseudobulge formation (e.g. Li and Shen 2012).



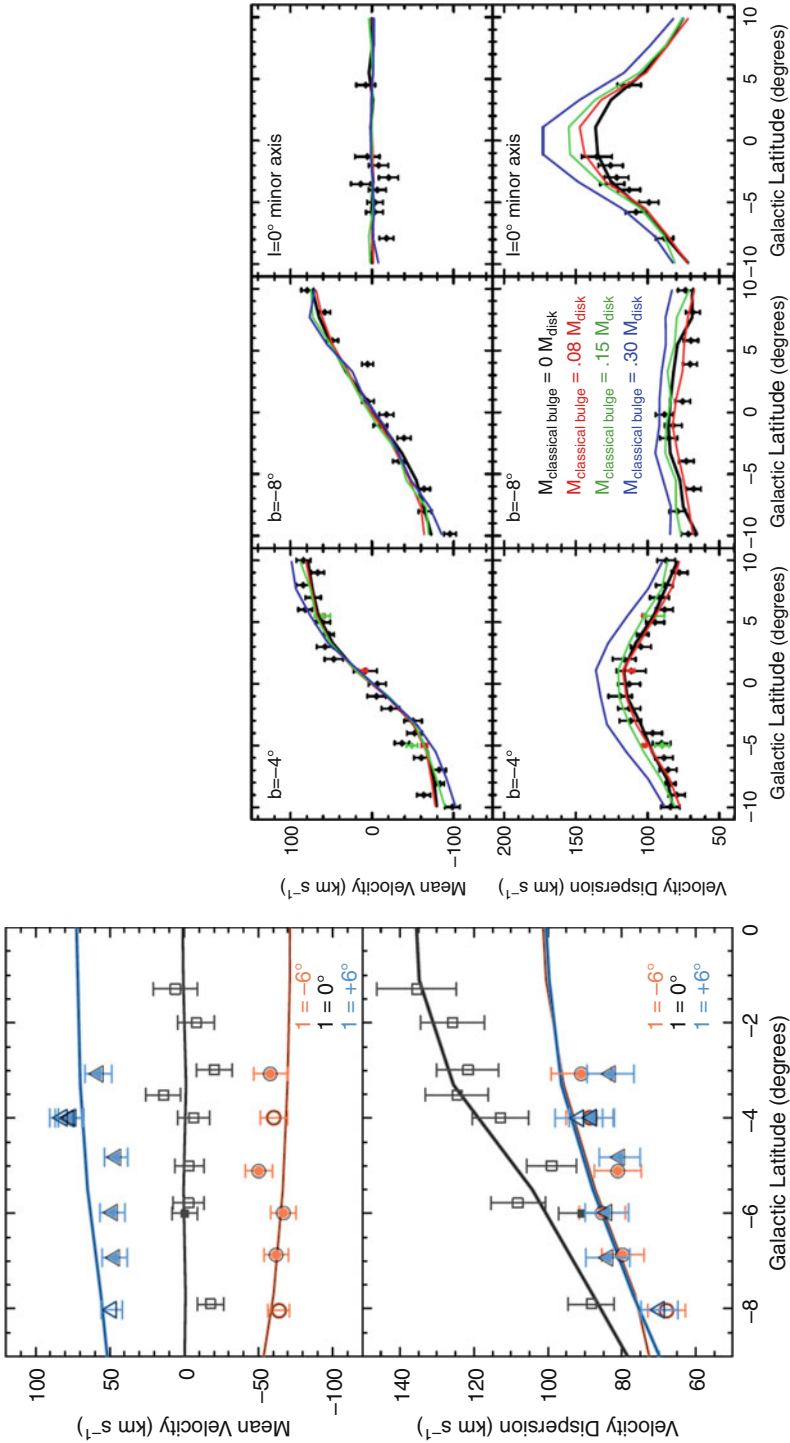
■ Fig. 6-32

(Left) *BRAVA* survey red giants (Kunder et al. 2012) following the sample selection defined by Rich et al. (2007a). The selection is designed to be unbiased with respect to metallicity and is cut at $K < 8$ to exclude foreground disk stars; the faint extension (blue points; $b = -8$) was required in order to have enough stars in the lower density high latitude fields. (Right) *BRAVA* fields identified by year of observation. The Southern galactic bulge was heavily sampled, avoiding the most reddened regions close to the plane. Roughly 100 stars were observed per field

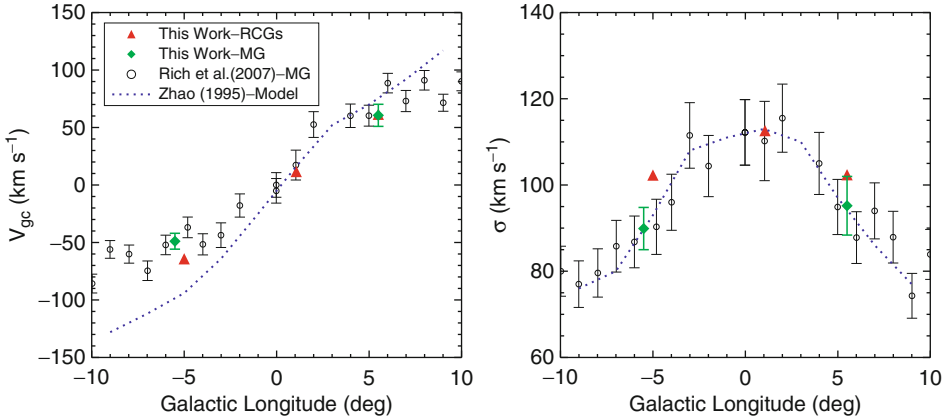


■ Fig. 6-33

Rotation and velocity dispersion profiles for bulge giants from the *BRAVA* survey (Rich et al. 2007a; Kunder et al. 2012). Models from Shen et al. (2010) are plotted as solid lines. The right-hand panel includes the Beaulieu et al. (2000) planetary nebula velocities. The self-consistent bulge model of Wang et al. (2012) also fits the data



■ Fig. 6-34 (Left) Minor-axis rotation and velocity dispersion profiles from BRAVA (Kunder et al. 2012). (Right) Constraints from Shen et al. (2010) that limit the fraction of a classical bulge to <10%



■ Fig. 6-35

The imaging Fabry–Perot survey of Rangwala et al. (2009) obtained radial velocities in three fields in agreement with the BRAVA dataset. The FP method builds a line profile of one Ca triplet line, and has the advantage that all detected stars can have radial velocity measurements. However, the velocities are based on the constraint of a single Ca line. The Fabry–Perot technique is to be exploited in new surveys using the SALT telescope in South Africa

The complete dataset for the BRAVA survey is maintained on the IRSA archive <http://irsa.ipac.caltech.edu/data/BRAVA> and at <http://brava.astro.ucla.edu/>.

The new radial velocity/abundance survey of Ness and Freeman (2012; ● Fig. 6-38) is less densely sampled than BRAVA but also extends 20° off the minor axis, into the disk. The rotation curve agrees with BRAVA, but the new data have $[\text{Fe}/\text{H}]$ and $[\alpha/\text{Fe}]$ measurements from low-resolution spectra. One surprising result is the apparent elevation of the alpha abundances over the entire inner 20° of longitude. If confirmed, one may interpret this alpha enhancement as reflecting the chemical evolution of the inner proto-disk that gave rise (dynamically) to the bar.

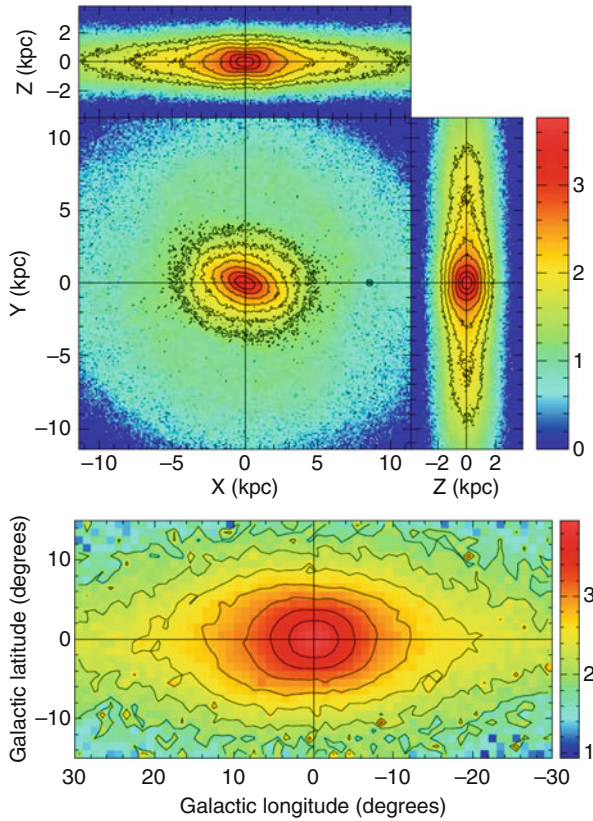
BRAVA did not find prominent cold structures in its survey region (● Fig. 6-37). The velocity distributions in each field are consistent with Gaussian, and no statistically significant cold streams or extreme velocity stars are noted.

Another approach to large-scale radial velocity surveys uses Fabry–Perot imaging of bulge fields in the Ca II 8542 absorption line (Rangwala et al. 2009). ● Figure 6-35 below shows that their results are in excellent agreement with the traditionally measured multiobject spectra. Imaging FP surveys are powerful in that they obtain a radial velocity measurement for every star in the field with the Ca triplet line, obviating the pitfalls of sample selection.

The imaging FP approach requires photometric conditions during the acquisition of the data, and calibration to a metallicity scale is in principle challenging. The combination of the two methods – perhaps with wider FP imaging and some spectroscopic observations within each FP field – might prove to be a powerful combination in the future.

4.2 Proper-Motion Studies

The velocity dispersion of the bulge corresponds to $\approx 2\text{--}3 \text{ mas year}^{-1}$ in proper motion, a value that is easily measurable with ground- and space-based techniques. A compilation of bulge proper-motion study results is given in Wang et al. (2012).

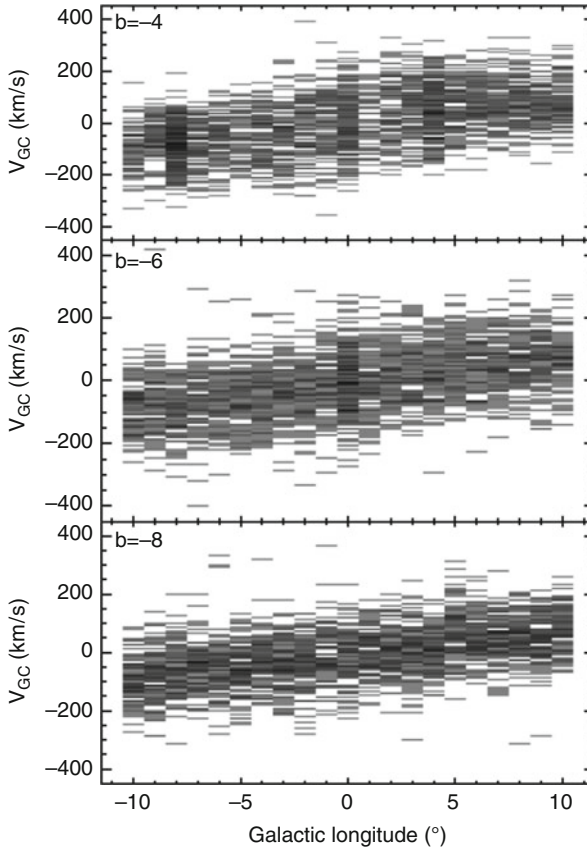


■ Fig. 6-36

Shen et al. (2010): *Upper three panels*: face-on and side-on views of the surface density of the best-fitting model N-body model. The Sun's position 8.5 kpc from the Galactic center is marked along the +x axis. The Galaxy rotates clockwise as seen in the face-on projection. *Bottom panel*: model surface brightness map in Galactic coordinates as seen from the Sun's location. Our perspective makes the box-shaped, edge-on bar appear to have greater vertical extent on the near side. The Galactic boxy bulge is observed to be similarly distorted

Proper-motion studies in the bulge have proven powerful as a means of separating populations (as in Kuijken and Rich 2002 and Clarkson et al. 2008; ▶ Figs. 6-9 and ▶ 6-10) and in exploring correlations between composition and kinematics (Zhao et al. 1996; Soto et al. 2007, 2012; ▶ Fig. 6-37). One of the most useful constructs is the velocity ellipsoid, in which the longitudinal proper motion is correlated with radial velocity. This method requires stars bright enough to obtain spectroscopy for radial velocities and abundances and is hence presently limited to giants. New integral field spectrographs will enable much larger samples of giants and even main sequence stars, to be studied. If such a correlation is present, it is an indication of bar kinematics. A new technique unveiled in Clarkson et al. (2008) is the use of photometric parallaxes. The correlation of proper motion with distance, as derived from isochrone fitting can yield constraints on the rotation.

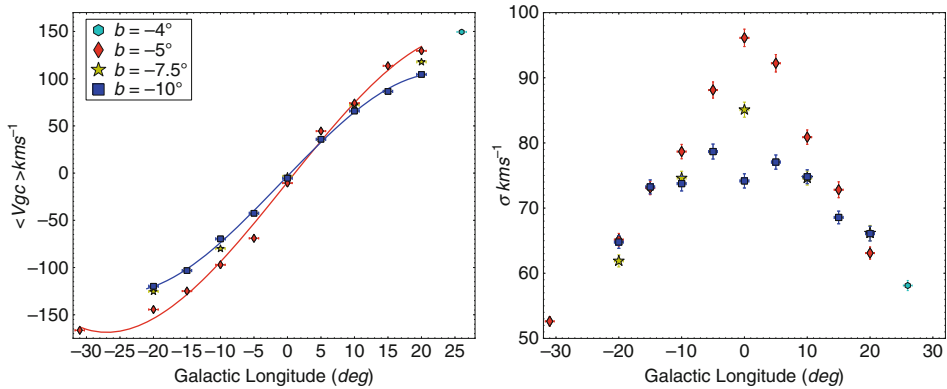
The earliest proper motions were based on historic photographic data (Spaenhauer et al. 1992). During the modern period, wide-field surveys have been provided by the OGLE project



■ Fig. 6-37

Bulge longitude velocity plots (Kunder et al. 2012) show no clear indication of either cold streams or a hotter, more slowly rotating “classical bulge” component. The cylindrical rotation is well illustrated here and is seen in all fields. In comparing with the gas kinematics, note that the bulge stars have large ($\sim 100 \text{ km s}^{-1}$) random motions; hence, the velocity field shows slower rotation and less structure compared with the gas. The K giants are also no closer to the plane than 500 pc (in contrast to stars in the OH/IR population)

(Sumi et al. 2004) and also via the application of historical plate material (Vieira et al. 2007) with new data from the OGLE III survey coming soon at the time of this writing. These proper motions have proven valuable in large sample abundance/kinematic correlations (e.g., Babusiaux et al. 2010). Rattenbury et al. (2007a) showed that the proper-motion dispersions follow the expected declines with increasing Galactic longitude and latitude, but noted significant variations between fields, and compared the dataset with made-to-measure models. The Southern Proper Motion Program (Girard et al. 2011; SPM4) covers the entire bulge, however its 2–8 mas accuracy, while sufficient to reject foreground stars (Johnson et al. 2012b), does not permit exploration of internal kinematics of the bulge. With a new generation of proper-motion data from OGLE III expected shortly, the promise of increased accuracy may allow the proper-motion data to constrain the orbit families responsible for sustaining the bar. The full potential of multiple HST datasets across the bulge remains to be exploited.



■ Fig. 6-38

Rotation and dispersion profiles as a function of Galactic latitude from the study of Ness and Freeman (2012); the results are similar to BRAVA but span twice the range in latitude. This study finds a low dependence of rotation on metallicity and confirms the cylindrical rotation of the BRAVA study

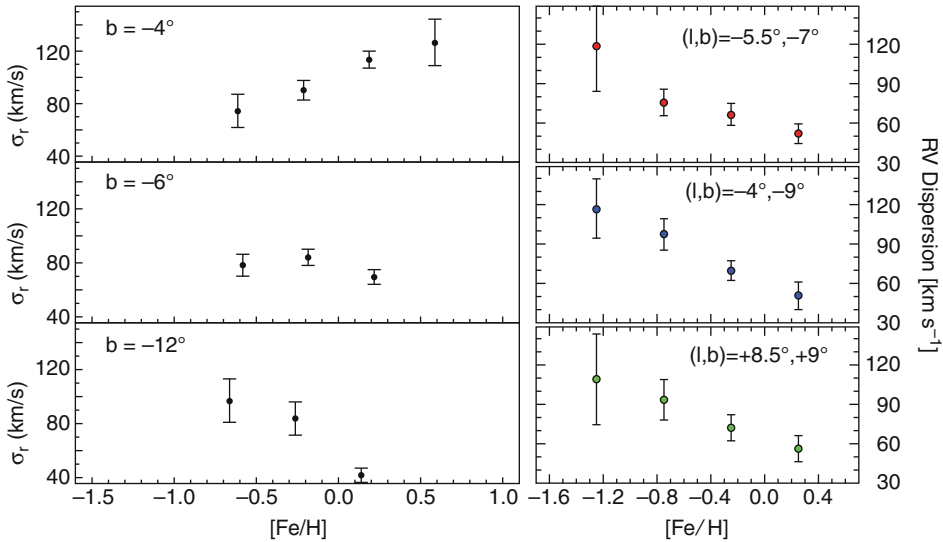
5 Kinematics and Composition

There is a strong expectation that composition and kinematics should be correlated, recalling the fundamental trends reported by Eggen et al. (1962), and conforming to the intuitive picture of collapse, dissipation, enrichment, and dissipative spin-up of gas. Rich (1990) found a correlation in the expected sense, with metal-rich stars having a lower velocity dispersion than metal-poor stars, a result that was confirmed by Sharples et al. (1990). Minniti (1996a) and Minniti et al. (1996) found rotation and dispersion to be strongly correlated with metallicity in the “expected” sense smaller velocity dispersion with increasing metallicity; the latter study considered a proper-motion-selected field near M22 at $(l, b) = (9.9^\circ, -7.6^\circ)$. Johnson et al. (2011) find these trends in the Plaut field at $b = -8^\circ$. Babusiaux et al. (2010) find the opposite: metallicity and velocity dispersion to be anticorrelated in Baade’s Window ($b = -4^\circ$). The problem is complicated.

One of the most intriguing correlations was found by Zhao et al. (1994): the most metal-rich stars exhibit a correlation between transverse proper motion and radial velocity, producing “vertex deviation” – a velocity ellipsoid with its major axis angled off of normal (► Fig. 6-40 below). Zhao et al. (1994), Soto et al. (2007) and Babusiaux et al. (2010) all find that this vertex deviation begins near $[\text{Fe}/\text{H}] \sim -0.5$. As the vertex deviation appears to be related to stars having orbital properties consistent with a bar (Zhao et al. 1996), this line of evidence suggests that most of the bar population has $[\text{Fe}/\text{H}] > -0.5$ and that one should be searching for kinematic/abundance transitions at $[\text{Fe}/\text{H}] < -0.5$.

5.1 Are There Subcomponents in the Bulge Abundance Distribution?

At the time of this writing, the observational landscape concerning subpopulations in the bulge is unsettled. Bensby et al. (2011) argue that the present abundance distribution of microlensed bulge dwarfs strongly favors distinct metal-rich and metal-poor populations, however, their

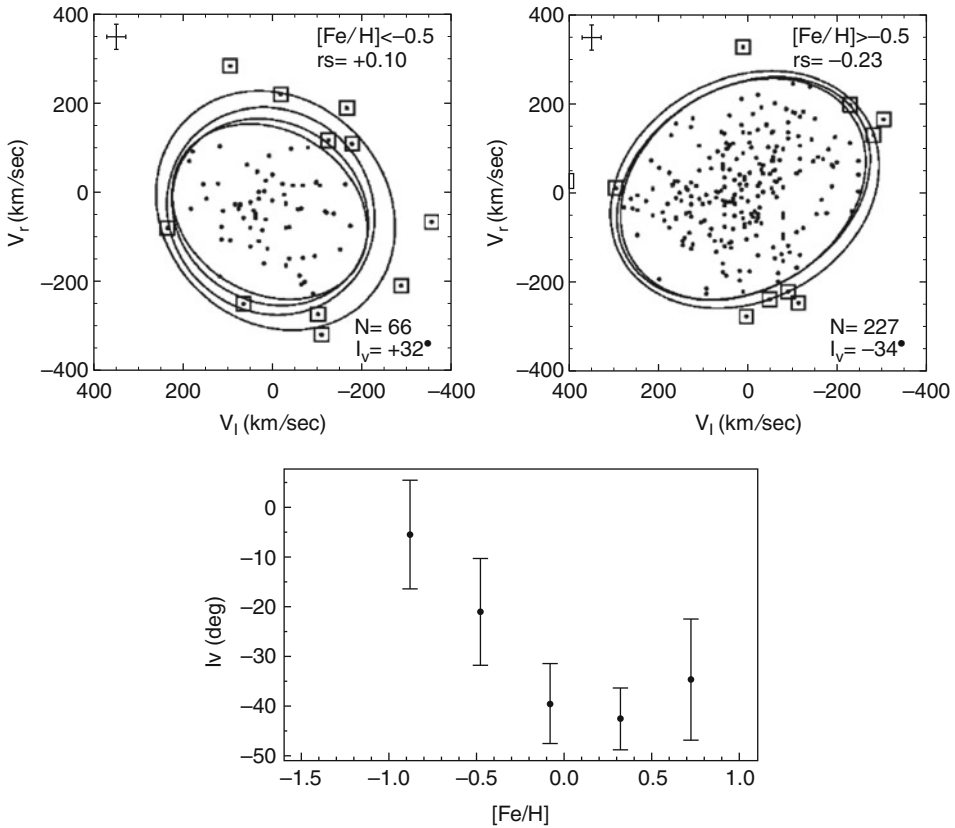


■ Fig. 6-39

Two views of kinematics and abundance in the Galactic bulge. Both studies employ multiobject high-resolution spectroscopy. (*Left*) Babusiaux et al. (2010) argue for a hot metal-rich population concentrated to the plane; all fields are on the minor axis. (*Right*) Johnson et al. (2012b) find abundances and kinematics to be correlated in the “expected” sense (e.g., Rich 1990; Minniti 1996a) of lower velocity dispersion with higher metallicity in off-axis bulge fields. Even though the sample sizes are 800 and 300 stars, respectively, it is clear that much larger samples covering more of the bulge are required to sort out which if any trend is correct

larger sample of 58 stars no longer shows the bimodality. Hill et al. (2011) deconvolve the generalized histogram of the Zoccali et al. (2008) abundance distribution in Baade’s Window into two subcomponents, both of which, however, would fall into the $[Fe/H] > -0.5$ regime where all of the stars evidently follow the bar-like vertex deviation. Ness et al. (Fig. 6-38) find evidence for subpopulations with amplitudes that vary as a function of Galactic latitude. Babusiaux et al. (2010) find that metallicity and velocity dispersion are *anticorrelated* at $b = -4^\circ$ and that these trends reverse themselves by -12° ; they argue that a kinematically hot stellar population of metal-rich stars is concentrated to the plane. Johnson et al. (2012b) do not confirm these trends in off-axis bulge fields (Fig. 6-40); they further find no statistical support for bimodality in any of their fields. Ness and Freeman (2012) argue that the trend of $[\alpha/Fe]$ versus $[Fe/H]$ can be subdivided into five populations by abundance and alpha enhancement, which then can be traced to vary spatially in the bulge as well (Fig. 6-41); the metal-rich subcomponent shows a strong concentration toward the plane, but the Babusiaux et al. (2010) kinematic trends are not confirmed. An object lesson here is that even samples of ~ 800 stars with high-dispersion spectroscopy are insufficient to settle definitively questions of subpopulations and trends between abundance and kinematics (Fig. 6-39).

The question of chemodynamical subpopulations will be best addressed with much larger samples than even the $\sim 10,000$ stars available at present. The correlation of composition and kinematics may prove valuable in validating the reality of subpopulations, but such efforts must await larger sample size and composition measurements of greater precision. It is likely that



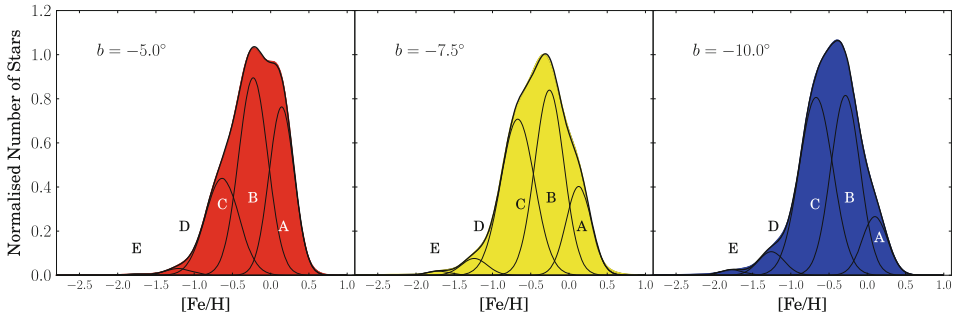
■ Fig. 6-40

Vertex deviation of the velocity ellipsoid, l -velocity derived via proper motion, versus radial velocity for individual stars. The vertex deviation is a strong signature of bar kinematics. Upper figures are from Soto et al. (2007); the lower figure plotting the vertex deviation in degrees versus $[\text{Fe}/\text{H}]$ is from Babusiaux et al. (2010). This technique will grow in impact as sample sizes increase

a multivariate approach (combining, e.g., the La/Eu ratios, Na/Al, with proper-motion measurements) over a range of Galactic latitudes might serve to solidify the emerging trends seen at present. Soto et al. (2007), Hill et al. (2011), and Ness and Freeman (2012) all stress the notion that a physical cause for bimodality should be sought in the underlying kinematics and abundances. It would appear that this is the best way forward.

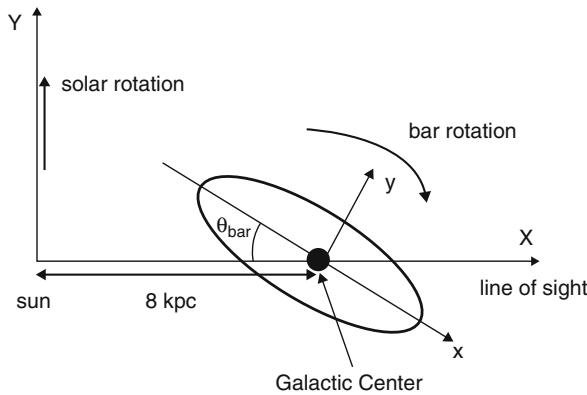
6 Structure

The Milky Way bulge appears to have three main structures that are common to many barred spiral galaxies. The main component is a boxy bar with a ~ 300 pc vertical scale height that likely hosts an X-shaped structure visible away from the plane. This is the bar that causes the asymmetric thickening at positive Galactic latitude on the COBE images (► Fig. 6-1); the geometry of the bar shown below in ► Fig. 6-42 pertains to this main bar (note that the derived



■ Fig. 6-41

A proposed approach to deconvolving the bulge abundance distribution into subcomponents based on $[\alpha/\text{Fe}]$ (Ness and Freeman 2012). The authors argue that the regular behavior of the amplitudes of these subpopulations with Galactic latitude and longitude argues for their reality, along with composition differences. Notice the strong concentration of subcomponent “A” to the plane; this is identified as the metal rich thin disk



■ Fig. 6-42

Coordinate system used to describe the bar geometry and its rotation. The near side of the bar is at positive Galactic longitude (from Wang et al. 2012). The angle $\theta_{\text{bar}} = 20\text{--}30^\circ$. The hypothetical long bar is at an angle $\sim 45^\circ$. The X-structure is perpendicular to the plane of the page

angle of this bar to the line of sight ranges from ~ 0 to 45°). This structure also contains most of the mass, roughly $2 \times 10^{10} M_\odot$. In terms of a Sersic index, the bulge is considered to be exponential (Sersic $n = 2$) and therefore satisfying one of the pseudobulge criteria and is not Sersic $n = 4$ (de Vaucouleurs spheroid). Although the X-shape has been found in external galaxies (Bureau and Athanassoula 2006) it is only recently that the X-shape has been reported and studied in the Milky Way. And it remains to be confirmed that the putative Milky Way X-structure is the same as that seen in the external galaxies.

The “long bar” is thin (~ 100 pc scale height) and lies in the plane and was discovered by infrared counts as well as the Spitzer GLIMPSE survey (Benjamin et al. 2005). Finally, there is a still debated nuclear bar or disk of extent ~ 100 pc, formerly known as the r^{-2} spheroid. The long and nuclear bars have similar masses, $\sim 10^9 M_\odot$. The classical bulge component has proven elusive to detect and quantify. That component (if shown to exist) would be a hotter,

more extended spheroidal population of old, metal-poor stars; if it has been spun up by the bar (Shen et al. 2010; Saha et al. 2012), it may be impossible to separate from the rest of the bulge population. [▶ Table 6-1](#) (Vanhollebeke et al. 2009) summarizes the properties of the bulge structures.

de Vaucouleurs (1964) argued from noncircular gas motions that the Milky Way might host a bar. Subsequent analyses using better data and models (e.g., Liszt and Burton 1980; Binney et al. 1991; Englmaier and Gerhard 1999; Fux 1999, see also [▶ Figs. 6-29](#) and [▶ 6-30](#)) strengthened the case for the bar potential. However, Blitz and Spergel (1991) used the balloon-borne IR survey of Matsumoto et al. (1982) and produced a model with the bar near the orientation that is widely adopted at present (inclined 25° to the line of sight and in the plane). This is the most prominent bar that is easily revealed via a range of observations.

Investigators rapidly sought to confirm the reality of the bar in actual star counts of sources. The same asymmetry in IRAS sources was noted by Nakada et al. (1991), in Mira variables (Whitelock and Catchpole 1992) and in AGB stars (Weinberg 1992). Dwek et al. (1995) easily detected the bar in the *COBE* infrared data, producing a number of models for the bar's structure. Stanek et al. (1997) used red clump giants to map the spatial distribution of the bar and to constrain the bar model, yielding a 3.5:1.5:1 axis ratio and a sightline angle from 20° to 30° . Babusiaux and Gilmore (2005) benefitted from wide-field infrared detectors and were able to overcome extinction issues; their work also yielded a tight constraint on the bar angle ($22^\circ \pm 5^\circ.5$). Deguchi et al. (2004) detect the bar in red giant SiO masers (their Fig. 6). Gonzalez et al. (2012) repeated the Babusiaux and Gilmore infrared approach, using the VVV dataset. [▶ Figure 6-43](#) illustrates some of these maps of the bar.

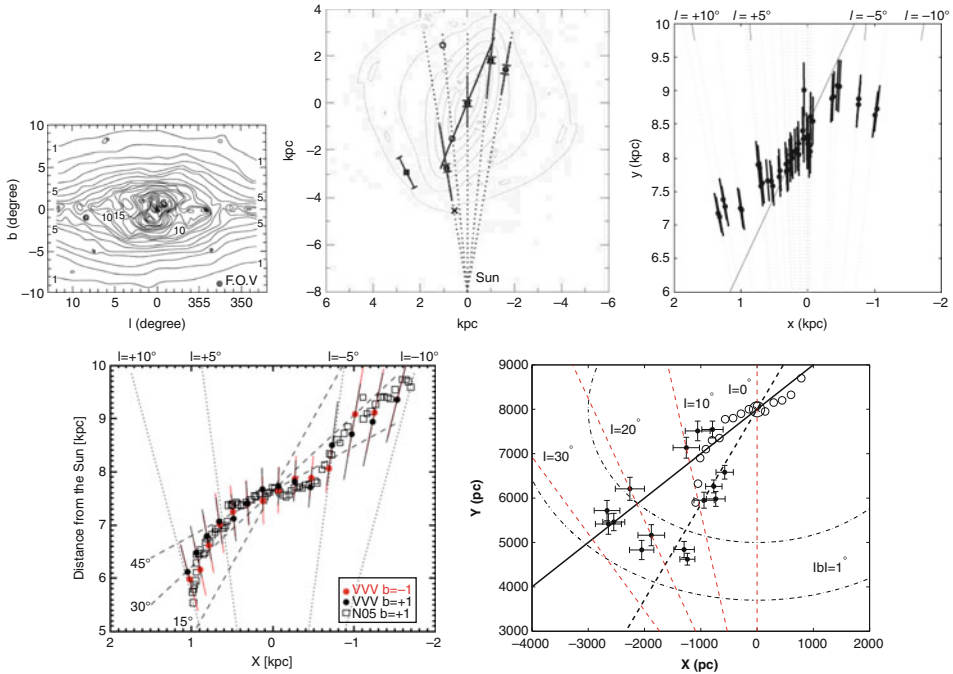
Additional early supporting evidence for bulge triaxiality came from modeling of the microlensing optical depth, which was rapidly understood to require a massive, triaxial structure (e.g., Zhao et al. 1994). The success of the Zhao (1996) self-consistent, rapidly rotating, triaxial bulge further supported the stable triaxial model. The bar has an approximately 1:(4–5):3 triaxial shape.

The main bar is actually $\sim 2\text{--}3$ kpc in radius, which projected on the sky would span from -14° to $+14^\circ$ in Galactic longitude if the bar were oriented perpendicular to our sightline. However, if the origin of the bar is in the buckling of the massive disk, then one might speculate that the “disk” population in the inner 2 kpc would share the properties of the bar, that is, being alpha-enhanced, metal-rich, and exhibiting a vertical abundance gradient. Some evidence that the inner disk and bar share a common chemical evolution history comes from Johnson et al. (2011) and as previously discussed, is also found in Ness et al. (2012). Red clump stars ~ 4 kpc from the nucleus are found to be metal-rich and alpha-enhanced, essentially indistinguishable from the Galactic bulge population.

The principal evidence supporting the long bar is the asymmetry in the disk and statistical analysis of Spitzer/GLIMPSE star counts (Benjamin et al. 2005). Hammersley et al. (2001), López-Corrodoira et al. (2007) and Cabrera-Lavers et al. (2007) argued for the long bar via infrared red clump star counts. The challenge of studying the long bar is that membership in it is statistical: a given star in that direction might belong to the disk, the bulge, or the long bar. Disentangling the populations will require proper motions and, likely, parallaxes. The long bar is also only detected in the first quadrant, and one would like to see a complete detection over the full length. Martínez-Valpuesta and Gerhard (2011; [▶ Fig. 6-44](#)) and Athanassoula (2012) argue that the long bar is not distinct and is part of the “main bar”; indeed, it is possible to trace the long bar only in the first quadrant and at an angle relative to the main bar that is not great enough for the overall structure to truly resemble multiple bar systems seen in other galaxies. Gerhard and Martínez-Valpuesta (2012) argue that two components with the same

■ **Table 6-1**
A not complete overview of recent values of parameters describing the GB, its bar and the distance to the GC (Source: Vanhollebeke et al. 2009)

Reference	R_0 (kpc)	a_m (kpc)	a_0 (pc)	$1-\eta, \zeta$	ϕ ($^\circ$)	Based on
Fernley et al. (1987)	8.0 ± 0.65					RR Lyrae stars
Reid et al. (1988)	7.1 ± 1.5					H ₂ O maser spots
Whitelock (1992)	9.1			1:0.25:0.25	45	Mira variables
Dwek et al. (1995)				$1:0.33 \pm 0.1$	20 ± 10	COBE/DIRBE surface brightness map
Binney et al. (1997)		1.9	100	1:0.6:0.4	20	COBE/DIRBE surface brightness map
Feast (1997)	8.1 ± 0.4					RR Lyrae stars
Stanek et al. (1997)				1:0.43:0.29	20–30	Red clump stars
Freudenreich (1998)		2.6				DIRBE full-sky surface brightness map
Paczynski and Stanek (1998)	8.4 ± 0.4					Red clump stars
Udalski (1998)	8.1 ± 0.15					RR Lyrae stars
Udalski (1998)	8.1 ± 0.06					Red clump stars
Sevenster et al. (1999)		2.5			44	OH/IR stars
Bissantz and Gerhard (2002)		2.8	100	1:(0.3–0.4):0.3	20–25	COBE/DIRBE L-band map
Eisenhauer et al. (2003)	7.94 ± 0.42					Stars orbiting black hole
Robin et al. (2003)				1:0.27:0.27	11.1 ± 0.7	Hipparcos data
Merrifield (2004)				1:0.6:0.4	25	H I gas and COBE/DIRBE surface brightness map
Babusiaux and Gilmore (2005)	7.7 ± 0.15				22 ± 5.5	Red clump stars
Eisenhauer et al. (2005)	7.62 ± 0.32					Stars orbiting black hole
Groenewegen and Blommaert (2005)	8.8 ± 0.4				47	Mira variables
López-Corrodoira et al. (2005)	$7.51 \pm 0.10 \pm 0.35$			1:0.5:0.4	20–35	2MASS star counts
Nishiyama et al. (2006)						Red clump stars
López-Corrodoira et al. (2007)					43	Red clump stars
Rattenbury et al. (2007b)				1:0.35:0.26	24–27	Red clump stars
Ghez et al. (2008)	8.0 ± 0.4					Stars orbiting black hole
Gillessen et al. (2009)	8.33 ± 0.35					Stars orbiting black hole



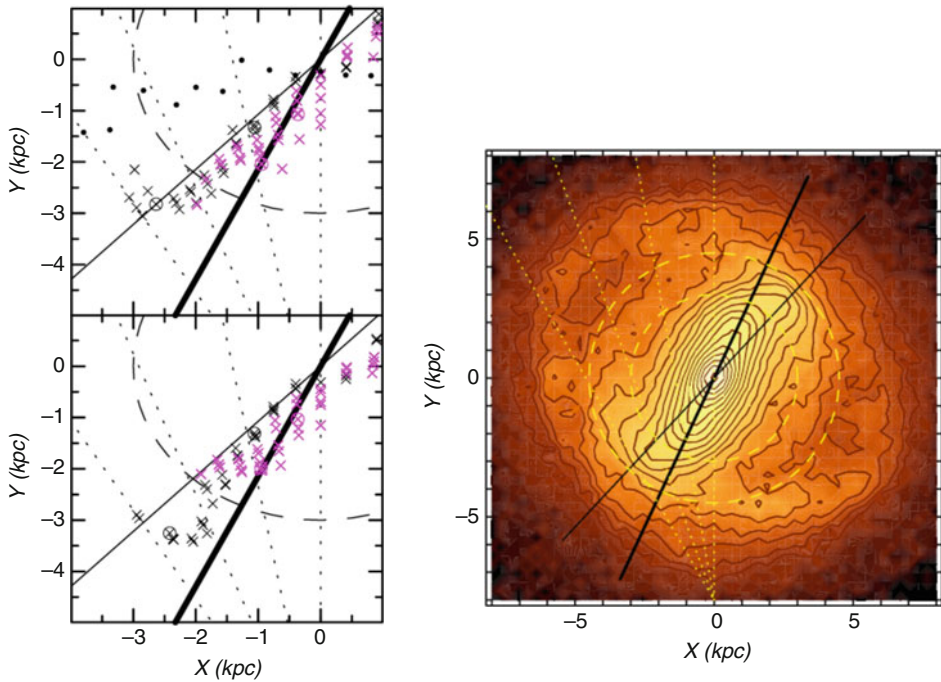
■ Fig. 6-43

A range of realizations of the bar. (*Upper left* projected light) Blitz and Spergel (1991) and Matsumoto et al. (1982): The asymmetric isophotes are consistent with one end of the bar being closer to the solar position. (*Upper middle*) Bar locus defined from the distance centroid of red clump stars using infrared imaging (Babusiaux and Gilmore 2005). (*Right*) Modeling of red clump star distances from the optical OGLE III survey (Rattenbury et al. 2007b). (*Lower left*) Red clump measurements used to define the bar in the VVV survey (Gonzalez et al. 2011c) (*Lower right*) Spatial distribution of red clump star distances at low Galactic latitude $|b| = 1^\circ$. There appears to be a second bar-like structure emerging at low Galactic latitude; this might be an independent long bar or part of the main bar (López-Corredoira et al. 2007)

position angle but different axis-ratios for isodensity contours can mimic two components with differing position angles. Both observations (Erwin and Sparke 2002) and simulations (Debatista and Shen 2007; Shen and Debattista 2009) have showed that the two bars in a double-barred galaxy generally has a length ratio of 0.1 to 0.2. In MW the lengths of the bulge bar and the possible long bar are quite close, making their possible coexistence dynamically puzzling.

An additional concern with the long bar as a distinct component is that the configuration with the two so closely aligned bars may not be stable. Such close alignments are not seen in nature (● Fig. 6-45). However, given its extent (8 kpc in total length) and flatness, the long bar is likely to be related to the disk and might be younger than the bulge, even if it dynamically is part of the main bar. Observations and time will tell.

If it does exist, definition of the nuclear bar is challenging due to the ~ 30 mag of visual extinction toward the Galactic center and the presence of the Galactic plane, star formation, etc. Alard (2001) argues for a nuclear bar based on excess $2 \mu\text{m}$ light. Launhardt et al. (2002) find a nuclear r^{-2} cluster and a disk of 230 pc radius, while Gerhard and Martinez-Valpuesta (2012) argue that star counts from the VVV data require no distinct nuclear bar, but might



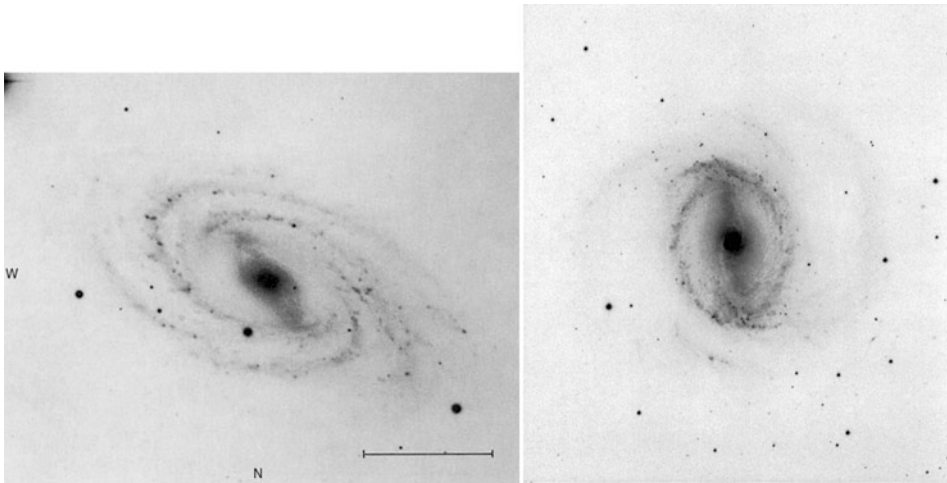
■ Fig. 6-44

(Left) Data and model that support the assertion that the long bar is actually part of the main bar. Location of the star count maxima in the Galactic plane, for fields near the disk plane (*black crosses*) and in the boxy bulge ($4^\circ < |b| < 8^\circ$, *pink crosses*). The *top panel* (a) shows the maxima for the model with leading curved ends of the bar; *black dots* show the maxima for the initial axisymmetric disk. The *lower panel* (b) is for a model with straight bar ends. The *thick solid line* shows the true orientation of the model, 25° . The *thin line* follows 43° . The *dashed circle* has a radius of 3 kpc. (Right) A single bar model can accommodate the main and “long” bars in a single structure (Martinez-Valpuesta and Gerhard 2011)

allow for a disk. The presence of the nuclear black hole and nuclear star cluster argues that this region is distinct in dynamics and, as previously mentioned, in stellar age distribution. It will probably be necessary to undertake a proper-motion/kinematic survey over the inner 100 pc if the issue is to be settled.

6.1 The X-Shaped Bulge

The 2MASS star counts revealed complexities in the red clump magnitude distribution (McWilliam and Zoccali 2010) that were confirmed by Nataf et al. (2010). Modeling of two peaks in the RC distribution, as a function of position, leads to the conclusion that they trace a component of stars with an X-shaped distribution (► Fig. 6-46); the spatial structure can also be seen clearly in the star counts of Saito et al. (2011). For latitudes $b < -5^\circ$, all the red clump stars appear to be members of the X-shaped structure (Saito et al. 2011). On the other hand, the bulge is observed to follow regular, cylindrical rotation in the same fields, and there is no evidence of subpopulations from the radial velocity surveys such as BRAVA.



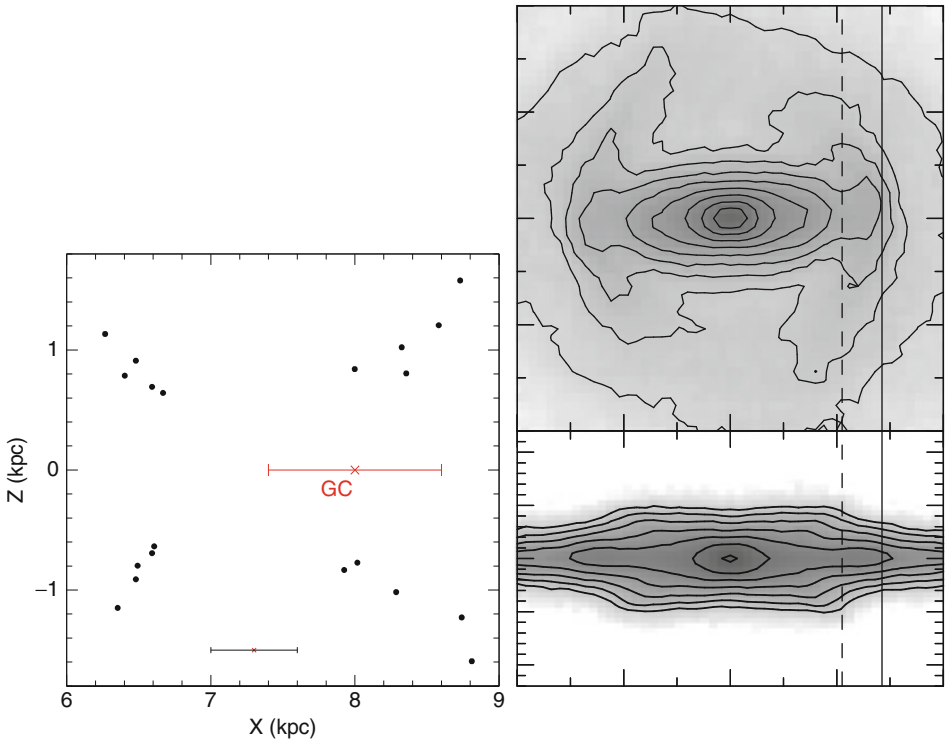
■ Fig. 6-45

External galaxies can exhibit misaligned bars, which may account for the “thick bar”/long bar dichotomy as opposed to the single bar in [Fig. 6-44](#). The geometry of the Milky Way’s “long” bar is somewhat different from these structures, as the thick and long bars are offset by only $\sim 15^\circ$ (Revised Shapley-Ames Catalog), NGC 3992 (*left*, Milky Way Revised Shapley-Ames Catalog), and NGC 1433 (*right*, Carnegie Atlas of Galaxies) http://ned.ipac.caltech.edu/level5/Shapley_Ames/RSA_contents.html

X-shaped bulges are well known in the extragalactic context and are associated with bars; they are also seen in N-body bar models (Athanasoula [2012](#); Li and Shen [2012](#); Ness et al. [2012](#)). It is not yet clear what fraction of mass in the bulge is accounted for by the X-shaped structure; at $b < -6^\circ$, Saito et al. ([2011](#)) argue that a substantial fraction of the red clump stars are in the X-structure, and refer to it as a dominant component within 2 kpc. However, it is difficult to definitively assign stars to the clump by the luminosity function alone. The strong likelihood is that the stellar population associated with the bar is responsible for the X distribution, as asserted by Ness et al. ([2012](#)).

At present membership of stars in an X-shaped bulge is still assessed in a statistical sense. de Propris et al. ([2011](#)) sought differences between the two clump populations as a function of abundances and kinematics and found none. It is possible that if the X-structure is associated with the bar, there will be little if any evidence of chemical or kinematic substructure within the bar, other than a lack of metal-poor stars ([Fig. 6-47](#)). However, Ness et al. ([2012](#)) have recovered the doubled clump from photometry associated with their large spectroscopically selected sample of metal rich stars. This work supports the notion that the X-structure is associated with the metal rich bar.

Li and Shen ([2012](#)) present clear evidence of an X-shaped structure in the Shen et al. ([2010](#)) bar/boxy bulge model. The X-shaped structure in their model is qualitatively consistent with the observed one in many aspects. The model X-shaped structure contains about 7% of light in the boxy bulge region, but it is significant enough to be identifiable in observations. An X-shaped structure naturally arises in the formation of bar/boxy bulges and is probably associated with orbits trapped around the vertically extended x_1 family. The X-shaped structure becomes increasingly symmetric about the disk plane, so Li and Shen ([2012](#)) conclude that the



■ Fig. 6-46

(Left) X-structure as mapped in red clump counts by McWilliam and Zoccali (2010). (Right) Density plot showing a model X-shaped bulge (face-on and edge-on projections) from Athanassoula (2005). A similar structure is seen in the model N-body bar model that fits the BRAVA dataset (Shen et al. 2010; Shen et al. 2012)



■ Fig. 6-47

The spiral galaxy NGC 4710 exhibits an X-shaped bulge as imaged by the Hubble Space Telescope and the Advanced Camera for Surveys. Credit: NASA, ESA, and P. Goudfrooj (STScI). More work will be required to show whether the X-shaped bulge claimed in the Milky Way is the same as similar structures observed in external galaxies like this one


observed symmetry may indicate that it formed at least a few billion years ago. There is no requirement that the X-shaped structure be comprised of a much younger stellar population.


6.2 A Classical Bulge?

Does there exist a slowly rotating population of old, metal poor stars that might be identified as a “classical” bulge? The *BRAVA* survey failed to detect such a population, even at the fringes of the *COBE* bulge light at $b = -8^\circ$. The survey strip turned up no evidence for a slowly rotating subcomponent nor is there evidence of a population transition at the “corners” of the bulge, $\pm 8^\circ$ (Howard et al. 2009). Ibata and Gilmore (1995) found both metal-rich stars and rotation at $b = -12^\circ$, not the expected signature for a classical bulge. Ness and Freeman (2012) segregate their K giant sample by metallicity, but only for $|l| > 10^\circ$ is the rotation observed to be slower for the metal-poor giants and, then, only by ~ 10 km/s. Rotation was also noted early, by Ibata and Gilmore (1995). No evidence of a kinematically distinct spheroidal population has clearly emerged.

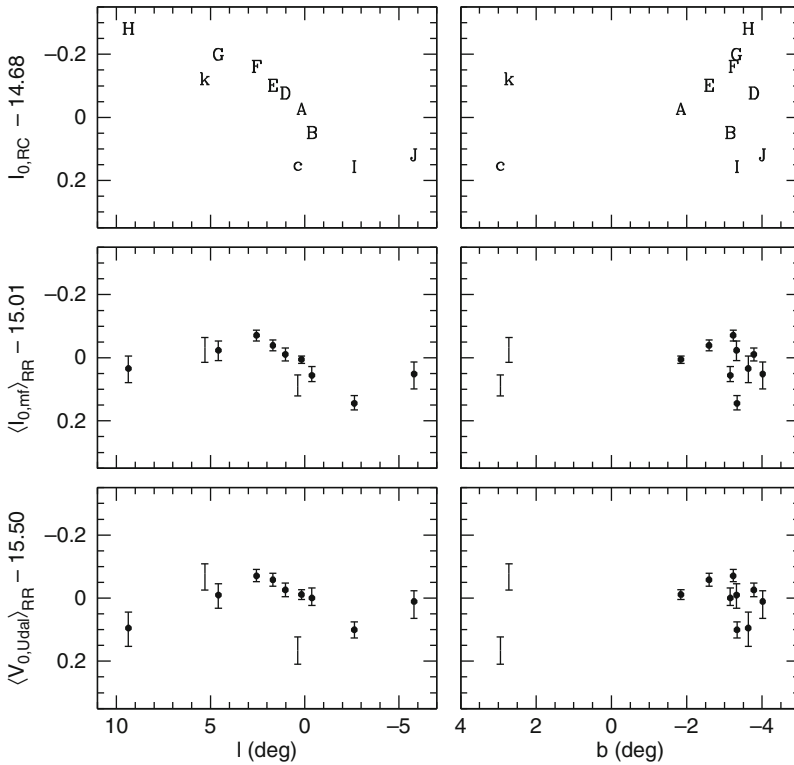
New theoretical simulations (Shen et al. 2010; Saha et al. 2012) show that a bar can spin up a preexisting classical bulge; the presence of a metallicity gradient in the outer bulge might indicate that this process is at work (note, however, that abundance gradients in peanut bulges are widespread; Kormendy and Illingworth 1982; Proctor et al. 2000). While the spun-up classical bulge might be indistinguishable from the bar at the present epoch, the models predict a considerable time lag, and the classical bulge must be in place before the spin-up can begin. Further, the models consider a $\sim 20\%$ by mass classical bulge component. Central to the model is the idea that the bar that originated in the disk has a very different origin from the classical bulge that it spins up. While the spinup would blur kinematic differences, the vastly different origins would likely be reflected in the composition trends, which would be reasonably expected to show a strong dependence on galactic latitude (as the “disky” bulge declined in density relative to the “classical” bulge). Observational tests will soon be in place to examine this.

The RR Lyrae population would be a good candidate to search for the classical bulge. Walker and Terndrup (1991) showed spectroscopically that these stars extend to low metallicity, a result that Kunder and Chaboyer (2008) confirm from their light curve analysis of 2,690 RR Lyrae stars observed in the course of the MACHO survey (full range from $-2.26 < [\text{Fe}/\text{H}] < -0.15$ and mean of $[\text{Fe}/\text{H}] = -1.25$). It is noteworthy that the mean metallicity of the RR Lyrae population is roughly at the tail of the bulge red giant distribution.

However, Collinge et al. (2006;  Fig. 6-48) use RR Lyrae stars as standard candles, showing that while the bar structure was present in the RR Lyrae population, it is clearly less prominent than for the red clump stars, but the bar is well detected.

The bulge globular cluster population might be proposed as a tracer of the classical bulge population. Minniti (1995a) argued that the “disk” system of globular clusters actually should be associated with the bulge. However, the bulge globular clusters extend to lower metallicity and also do not reach supersolar metallicity, compared with the giants ( Fig. 6-18). From a chemical evolution consideration, they would appear to be distinct from the bulge population and may well have the same origin as the bulge RR Lyrae stars.

Burkert and Smith (1997) examined the kinematics of the metal-rich clusters and find that $\sim 1/3$ of them may be associated with the bar. The most massive clusters rotate yet are centrally condensed; another subset of these clusters appears to exhibit the rapid rotation characteristic of the disk. The globular cluster system appears to be a poor candidate for tracing a putative



■ Fig. 6-48

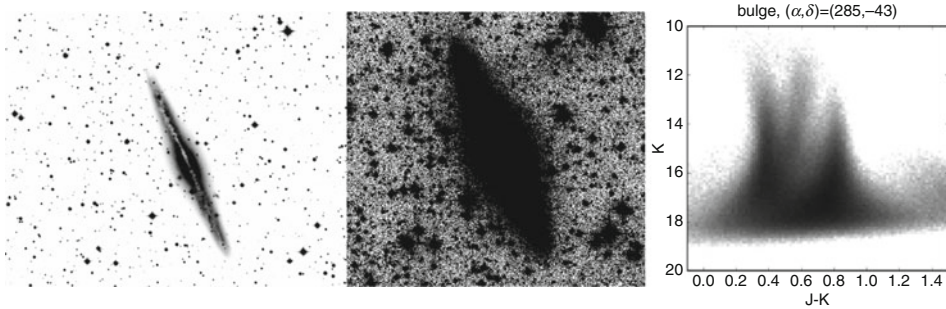
Collinge et al. 2006: Bar structure and distances on the major and minor axes for red clump giants (*upper panels*) and RR Lyrae stars (*middle, lower panels*). The *upper and middle panels* are I band while the lowest is V band. The letter designations refer to the fields of Sumi et al. (2004); lower case letters are fields at positive Galactic latitude. Notice that the bar tilt is much more prominent for the red clump stars than for the RR Lyrae stars which have mostly metal poor progenitors. Nonetheless, the bar remains clearly traceable for the RR Lyraes

classical bulge. In extragalactic systems, there is a red/blue cluster dichotomy, with red clusters being associated with the bulge and blue with the stellar halo.

The bulge RR Lyrae stars and globular clusters are two distinct populations where age and relative age can be securely dated and appear to have a spatial distribution that is not clearly coincident with the main bar. This may point to the main bar being younger than those populations. ☉ *Figure 6-49* suggests that there may be cause to search for a classical bulge population at higher Galactic latitude and wisdom in pushing composition/kinematic observations there. GAIA will provide superb astrometric data at high latitude and may well settle the issue.

7 The Milky Way Bulge in an Extragalactic Context

From a standpoint of considering the Milky Way bulge in the context of spheroids, it is critical to remember that the boxy bar with a likely X-shaped structure is observed in other galaxies only to exist in the context of a disk. There are no known isolated boxy “peanut” or “X” bars



■ Fig. 6-49

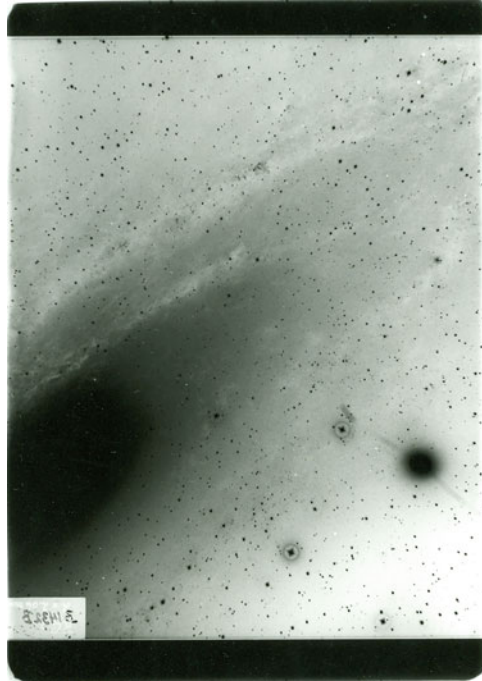
The Milky Way analog NGC 891 has a peanut-shaped bulge (*left*), but a deep image reveals what appears to be a thick disk and spheroid, perhaps even a “classical” bulge, although some of the central spheroidal population may be accreted (Mouhcine et al. 2010). Note that the full extent of the *BRAVA* survey (and the COBE bulge) is equivalently contained in the bulge of the *left-hand image*. The *right-hand panel* is a VISTA infrared color-magnitude diagram (Courtesy of G. Gilmore) at $b = -43^\circ$ that includes a red giant branch (*middle branch*) that would appear to correspond to a similar extended spheroidal population in the Milky Way. If present, such a “classical bulge” or inner halo accounts for very little mass and is likely difficult to detect in the study of the main bar (e.g., *left-hand panel*). If the Milky Way has such an extended population, it would not be physically possible for the bar to spin it up. The definition of this population is a task for the next decade. Images of NGC 891 are from the 0.7 m Saturn Lodge telescope at the Polaris Observatory Association (Rich et al. 2013). An extended bulge structure like this is found for a number of edge-on spiral galaxies

lacking disks. There are two examples of isolated S0 galaxies with boxy halos, but the inner high surface brightness regions exhibit clear S0 morphology (Tal et al. 2009; Graham et al. 2012). The bulge is too faint and not sufficiently strong-lined to be truly representative of the stellar populations inhabiting giant elliptical galaxies. Bars are common, however: roughly one third of spirals are barred in visible light (Sellwood and Wilkinson 1993) rising to 60% or greater in the infrared (Menéndez-Delmestre et al. 2007).

The first modern effort to consider the total luminosity of the spheroidal component was that of de Vaucouleurs and Pence (1978). Utilizing star counts and adopting the density law of the distribution of globular clusters, they found $R_e = 2.67$ kpc, $M_B = -18.67$, and $c/a = 0.6$. The large value of effective radius is not supported by modern studies that find a bulge scale height of ≈ 400 pc (e.g., Dwek et al. 1995). Tremaine et al. (2002) consider the range of infrared observations and arrive at $R_e = 700$ pc (► Fig. 6-50).

Kormendy et al. (2010) adopt a (pseudo)bulge/total = 0.19 for the Milky Way, based on Kent et al. (1991) and Dwek et al. (1995). The implied $M_K = -21.9$, and $M_V = -19.0$, placing our bulge ~ 2 mag fainter than L^* (the Milky Way bulge is not a luminous spheroid). Tremaine et al. (2002) classify the Milky Way as SBbc and give $M_B = -17.65$. The total infrared bolometric luminosity was derived by Dwek et al. (1995) to be $L_{\text{bol}} = 5.3 \times 10^9 L_\odot$. Kent (1992) found a dynamical mass of $2 \times 10^{10} M_\odot$; M/L_K has been derived to be ≈ 1 (see, e.g., Zoccali et al. 2000).

Whitford (1978) used a photometric scanner (recall ► Fig. 6-6) to demonstrate that the integrated light of Galactic bulge fields near Baade’s Window have the line strength typical of bulge regions 600 pc from the plane in well-known Sb galaxies like NGC 4565 – that it is a

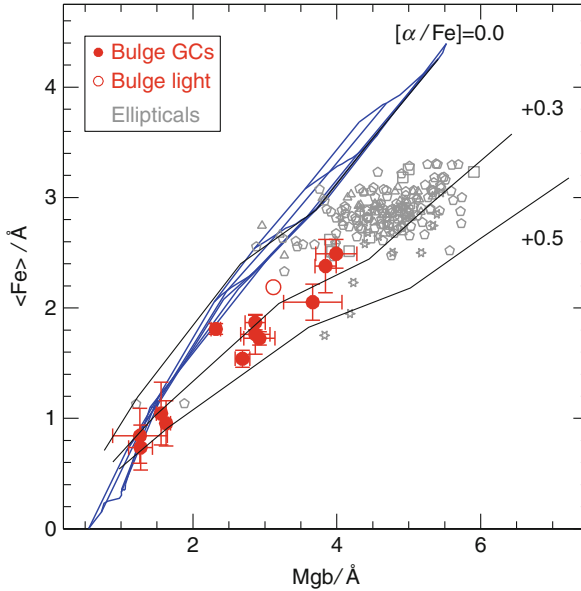


■ Fig. 6-50

The incipient resolution of the bulge of M31 into stars and its broad comparison with the population of the Galactic bulge in Baade's Window was one of the key advances that led to the concept of two stellar populations. This is a reproduction of Baade's original M31 resolution plate (Baade 1944) (Eastman 103aE plate behind Schott RG2 filter; Mt. Wilson 100-in telescope, 1942). Print courtesy of S. Majewski

“typical” bulge population. The work required a very careful correction for the foreground light of the disk, which remains the limiting factor in such measurements.

Using medium spectral resolution, Puzia et al. (2002) and Maraston et al. (2003: ▶ Fig. 6-51) compare the integrated light of bulge fields near Baade's Window with globular clusters and find a remarkable similarity between the bulge and the globular clusters NGC 6553 and 6528. It may be noted that even with careful avoidance of bright foreground dwarfs, the bulge field observations are not perfectly corrected for disk contamination due to stars near the turnoff and in the main sequence (note ▶ Fig. 6-9, this chapter) nor do the field observations properly account for hot horizontal branch and blue straggler stars, which are rare but present. However, one may safely conclude that the bulge shares the alpha enhancement of the giant ellipticals but is not as metal-rich as those populations. Future work on the integrated spectrum of bulge light might explore the disk subtraction technique of Zoccali et al. (2003) or might use proper-motion measurements to veto nonmember spectra obtained using an integral field unit. It is, however, unlikely that such corrections would place the bulge among the ellipticals as it is an order of magnitude less luminous than the typical bright elliptical galaxy.



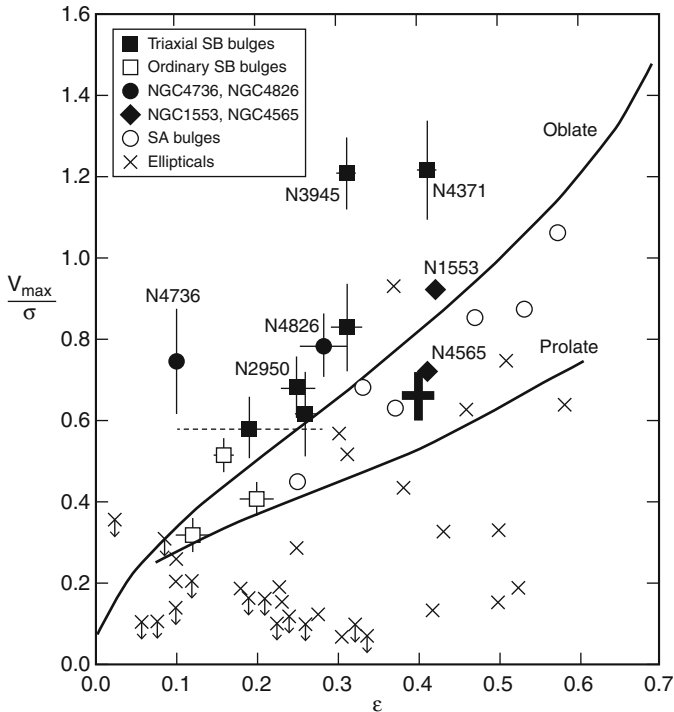
■ Fig. 6-51

Lick indices measured from the integrated light of globular clusters and bulge fields. The Galactic bulge falls on alpha-enhanced tracks but at lower metallicity than most elliptical galaxies (Maraston et al. 2003)

The velocity dispersion that would be observed for the Milky Way bulge in an extragalactic perspective has been carefully considered by Tremaine et al. (2002); see also [Fig. 6-31](#). They find $\sigma = 103$ km/s; it is noteworthy that the best estimate of the mass of the Milky Way central black hole is now $\sim 4 \times 10^6 M_{\odot}$, which places the Milky Way squarely on the $M - \sigma$ relationship for bulges.

The new rotation measurements of the *BRAVA* project now solidly confirm that the bulge falls near the oblate rotator line in the Binney diagram ([Fig. 6-52](#)) close to the boxy bulge of NGC 4565 and other pseudobulges. This is not a new result; Minniti et al. (1996) found bulge rotation from K giants and showed that the bulge falls near the oblate rotator line. Kormendy and Illingworth (1982) also noted that the boxy bulge of NGC 4565 exhibited cylindrical rotation and commented on the association of boxy bulges and cylindrical rotation. The Milky Way bulge beautifully conforms. Both NGC 4565 and the Milky Way bulges also exhibit vertical abundance gradients (Proctor et al. 2000).

Without repeating the discussion of Kormendy and Kennicutt (2004), it is difficult to assess to what extent the bulge of the Milky Way is a true pseudobulge. The classical pseudobulge examples (e.g., those in Kormendy et al. (2010) like NGC 6946) exhibit a nucleus and central brightening in excess of that expected from the disk but host significant dust, gas, and star formation. In the Milky Way, there exists a prominent bar with light dominated by old stars. While star formation is present in the central few tens of pc, it is not predominant over the bulk of the bar. One may conclude that the Milky Way is more similar to NGC 4565 rather than to the pseudobulge cases that appear to lack even an old, massive bar. Even allowing for



■ Fig. 6-52


Placement of the Galactic bulge in the Binney diagram. Notice that the bulge falls near the well-studied edge-on spiral NGC 4565 but does not exhibit the extreme rotation of the pseudobulge NGC 4736. The bulge is plotted as a solid cross, over the version of the Binney diagram from Kormendy and Kennicutt (2004)

the alpha enhancement and evidence that the bulge population is predominant, the $n = 2$ Sersic index and bar structure qualify the Milky Way for meeting one of the Kormendy and Kennicutt (2004) criteria for pseudobulges and therefore the Milky Way hosts a pseudobulge based on this classification scheme.

By defining virtue of its morphology, the Milky Way bulge does not meet the concept of Renzini (1999) as an “elliptical galaxy residing in a disk.” The population is far from a perfect analog of the giant ellipticals; our bulge is ten times less luminous, less metal-rich, has a Sersic index of 2 and not 4, and has a boxy (likely X) structure that is *never observed in isolation without a disk being present*. This latter issue is important; the class of boxy bulges and bars are always associated in nature with a disk. There is also evidence for continuous star formation in the central 100 pc as well, and our view to the Galactic plane is obscured and complicated by intervening stellar populations.

The bulge of M31 is ~ 100 times more distant but has also been well studied. There are many similarities to the Milky Way, even to the point of exhibiting boxy isophotes (Beaton et al. 2007). Kent (1989) modeled extent kinematics of the M31 bulge using an isotropic oblate rotator model, finding a mass twice that of the Milky Way bulge, $4 \times 10^{10} M_{\odot}$; note that

Saglia et al. (2010) call for a mass perhaps a factor of 2 greater. The more luminous, regular M31 bulge may therefore be a better match to the spheroidal populations of giant elliptical galaxies than the Milky Way bulge.

The M31 bulge can be resolved even from the ground to individual stars, which are shown to be M giants exhibiting a red giant branch similar to the bulge but extended to $M_{\text{bol}} = -5.5$ (Rich and Mould 1991) and dominated by M giants, like the Milky Way bulge (Rich et al. 1989). Using HST to resolve the red giant branch in the infrared, Stephens et al. (2003) found no compelling evidence that the M31 bulge differs from that of the Milky Way in its bolometric luminosity function or age. The abundance distribution resembles closely that of the Milky Way bulge (Sarajedini and Jablonka 2005;  Fig. 6-53).

Ground-based adaptive optics imaging of the M31 bulge by Olsen et al. (2006) measure no difference between bulge and disk fields, finding the bulge population to be older than 6 Gyr and metal-rich. It is unlikely that the bolometric luminosity function of M giants can discriminate between 5- and 12 Gyr-old populations at these high metallicities. However, Saglia et al. (2010) use spectroscopy to argue that the M31 bulge is old (12 Gyr) except for the inner few arcsec. It is also interesting that Lauer et al. (2012) find a cluster of blue stars with age 100–200 Myr surrounding the M31 nuclear black hole. It is possible that the most central regions of the M31 bulge contain a wider range of stellar ages as is the case for the Milky Way. Additional support for a wide range of ages in the inner parts of M31 is found in the halo field, which contains a 6–8 Gyr-old population only 11 kpc from the nucleus, even in locations hosting metal-rich stars (Brown et al. 2003, 2006). For the halo fields, this result was established by deep imaging to the main-sequence turnoff, something an approach that is not yet possible in the far more crowded bulge of M31.

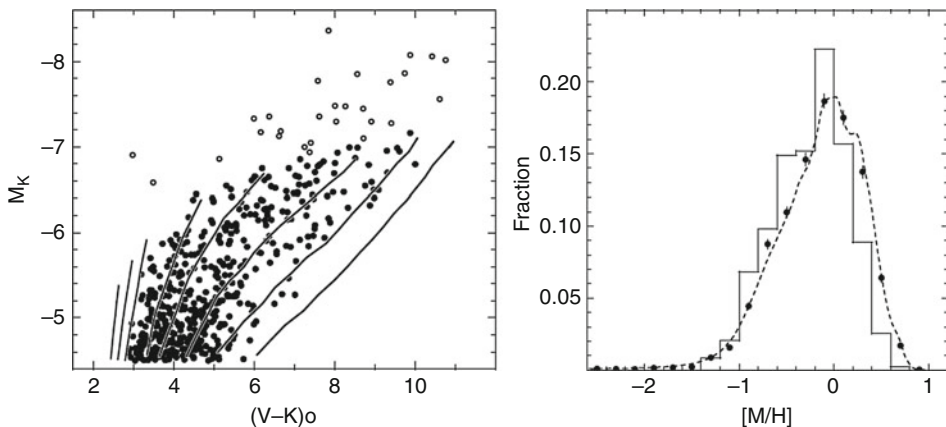


 Fig. 6-53


M31 bulge abundance distribution from HST imaging (Sarajedini and Jablonka 2005). (Left) Photometric metallicity derivation from $(V-K)_0$ color. (Right) Metallicity distribution of field stars in the M31 bulge derived via their photometric metallicities (filled circles) compared with the photometric metallicity distribution of the Galactic bulge fields, from Zoccali et al. (2003). The similarity is striking; note the absence of metal-poor stars in both populations

8 Theories for the Formation of the Bulge

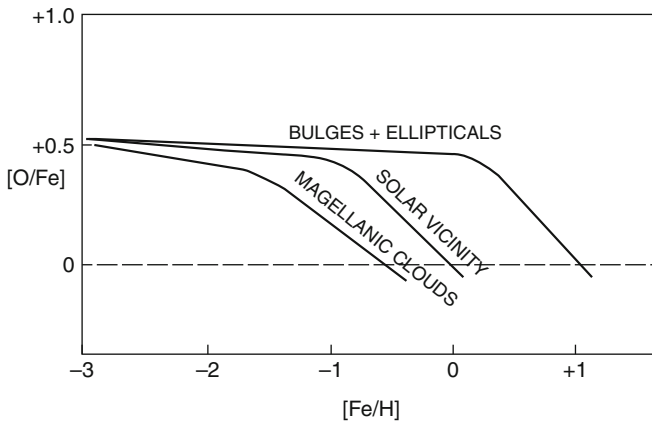
One contemporary theory of spheroid formation may be loosely described as an updated Eggen et al. (1962) scenario, with the merger-driven LCDM theory overlaid. Dark matter clumps (presumably carrying baryons) are accreted (e.g., Elmegreen 1999; Abadi et al. 2003). As pointed out by Kormendy et al. (2010), the dominance of the boxy bulge/bar and absence of a “classical” bulge component argue strongly against this scenario being correct for the Milky Way bulge. A variant of these ideas (Immeli et al. 2004) has the bulge forming from the merger of multiple, chemically distinct clumps.

The more widely accepted class of models involves the dynamical secular evolution of a massive disk that buckles into a bar (Combes et al. 1990; Raha et al. 1991; Norman et al. 1996; Shen et al. 2010). This class of models can produce a peanut-shaped bar and even X-shaped structures. Purely dynamical evolution can account for the structure and kinematics of the bulge that are observed today. It is also known that bar structures are long lived and can be observed to be in place even at redshift 1 (Sheth et al. 2008), and if the bulge is as old as the globular clusters, it would imply a formation redshift of 2–3. However, the observed abundance gradient in the Galactic bulge (and that of NGC 4565) cannot be simply reproduced by a dynamical process; one cannot construct a means to selectively accelerate stars of a given metallicity.

Bekki and Tsujimoto (2011) in fact confirm that the dynamical buckling of a disk cannot impose a metallicity gradient. They propose a two-stage formation process involving a thick disk and younger thin disk that would both produce the metallicity gradient and be consistent with the younger population suggested by the microlensing observations. They argue that the similarity of the bulge and thick disk populations (e.g., Alves-Brito et al. 2010) supports this scenario. The *BRAVA* results (Howard et al. 2009; Kunder et al. 2012) are inconsistent with a two-component model, however. The kinematic unity of the bulge is striking, and additional supporting evidence would be required to solidly support such a two-stage formation scenario.

Considering chemical evolution, the operating idea is that the bulge formed early and rapidly (Matteucci and Brocato 1990; Matteucci et al. 1999; Ballero et al. 2007; Matteucci et al. 2011) with early contributions from core-collapse SNe in <1 Gyr. The formation of the bulge would have required a $\sim 20 M_{\odot}$ star formation rate over 1 Gyr. The schematic result of such a rapid star formation is shown in  Fig. 6-54: the alpha elements are enhanced at low metallicity relative to the thin disk and might be enhanced even at greater than solar metallicity. The chemical evolution with rapid formation gave rise to efforts to fit the abundance distribution (Rich 1990) with the one-zone “simple model” of chemical evolution; the result was actually satisfactory in that the bulge abundance distribution has enough metal-poor stars to escape the “G dwarf problem.” Ballero et al. (2007) and Cescutti and Matteucci (2011) fit the metallicity distribution and run of $[\alpha/\text{Fe}]$ versus $[\text{Fe}/\text{H}]$, requiring a <1 Gyr formation timescale for the bulge and in the latter paper, favoring the Salpeter mass function. Given the complexity of the bulge’s composition trends, kinematics, and likelihood of some subpopulations, the simple closed-box one-zone model of chemical evolution can likely be ruled out.

The lack of a radial abundance gradient (Ness and Freeman 2012; Johnson et al. 2012b) along with the presence of a strong vertical abundance gradient, and kinematic disk origin, might be a point in favor of the notion that the formation of the massive disk was accompanied by a supernova-driven wind. Such processes are observed in present-day star-forming galaxies, most notably, in M82. Hartwick (1976) emphasizes that wind-driven loss of metals can reduce the effective yield in the Simple Model of chemical evolution (see also Mould 1984), and Dekel and Woo (2003) appeal to wind-driven feedback to explain the scaling relations of dwarf galaxies. The massive disk defines the potential well, which in turn sets the effectiveness of winds in



■ Fig. 6-54

Schematic trend illustrating chemical evolution of the Galactic bulge (Matteucci and Brocato 1990). The modern version would plot $[\alpha/\text{Fe}]$ on the y-axis. The rapid enrichment of bulges/ellipticals compared to the thin disk results in enhancement of alpha elements, even at suprasolar metallicity. The actual bulge trend of $[\text{O}/\text{Fe}]$ vs $[\text{Fe}/\text{H}]$ more closely resembles that of the “Solar vicinity” line (Bensby et al. 2012; Johnson et al. 2013 in prep.) although other alpha elements (Mg) follow a trend closer to the “bulges & Ellipticals” line

removing metals, naturally forcing the abundance gradient. It is attractive to consider winds as being a central factor in the early chemical evolution of the bulge, as winds appear to be ubiquitous in the formation of galaxies (even at $z \sim 5$, one may observe the wide dispersal of metals in the intergalactic medium). The connection of the chemical and secular evolution of the central bulge/bar system offers a very interesting line of investigation for the next decade.

There are already strong indications that no simple chemical evolution model suffices to describe the evolution of the bulge. There instead appears to be complex correlations between kinematics and abundances that will require chemodynamical models to address.

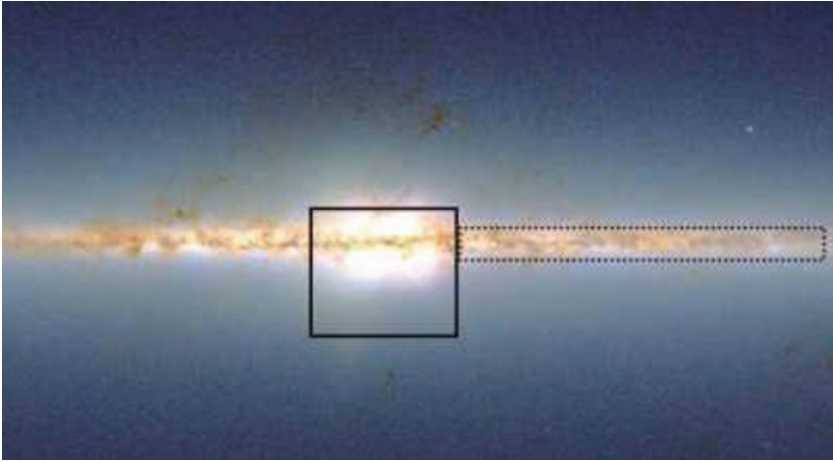
The strongest constraints on the formation of the bulge must await larger samples of stars with composition and kinematic measurements, especially proper motions. We are on the verge of the sample sizes and data quality needed to attain this goal.

9 Future Surveys

This review chapter is likely to become outdated within months of its printing, as many powerful surveys are coming online and releasing datasets in the next few years. By 2020, the community will be in possession of datasets so powerful that they will be able to tightly constrain bulge formation theories.

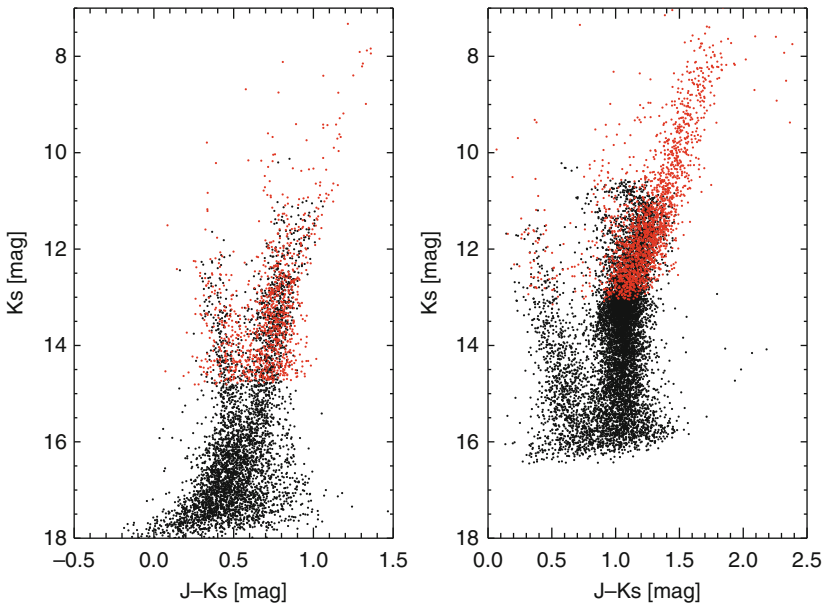
9.1 Ground-Based Imaging Surveys

The VVV survey (Minniti et al. 2010; ▶ Figs. 6-55 and 6-56) employs the 4 m VISTA infrared telescope and will cover 520 deg² from 2010–2014. The survey will have five bands ($ZYJHK_s$)



■ Fig. 6-55

2MASS map of the inner Milky Way showing the VVV bulge (*solid box*, $-10^\circ < l < +10^\circ$ and $-10^\circ < b < +5^\circ$) and plane survey areas (*dotted box*, $-65^\circ < l < -10^\circ$ and $-2^\circ < b < +2^\circ$) (Figure from Minniti et al. (2010))



■ Fig. 6-56

Comparison of the depth of color-magnitude diagram achieved by 2MASS (*red*) and VVV (*black*) for two bulge fields with very different crowding. Field b235 (Left) is near $(l, b) = (0, -7)$; Field b305 (right) is near $(l, b) = (0, -2)$. VVV uses a 4 m telescope with seeing at Paranal, hence can reach over 2 mag fainter (Saito et al. 2012)

spanning 0.9–2.5 μm ; multiple cadences of visits will be designed to yield light curves of variable stars including RR Lyraes, which could in principle be used to produce a 3D map of the bulge. The approximate depths (modulo source confusion) are $Z = 21.9$, $Y = 21.2$, $J = 20.2$, $H = 18.2$, and $K_s = 18.1$. The full survey is described in Saito et al. (2012). The survey has already produced reddening and metallicity maps and a map of the inner bar. Owing to its greater aperture and superior detectors, VVV already reaches 2 mag fainter than the 2MASS survey (● Fig. 6-53).

The bulge will also be imaged by the SkyMapper survey, which will have a six color *uvgriz* survey, replicating the Sloan Digital Sky Survey but reaching deeper in the u band than the Sloan survey. The 1.3 m telescope will operate in a stare mode, covering much of the southern sky before LSST. The filter set should enable selection of ultra metal-poor stars and modeling of the stellar population in the bulge.

The Dark Energy Camera (DECam) on the CTIO 4 m has a 3° field of view and employs the SDSS *ugriz* bands. The SDSS u photometry should be valuable in sorting out hot horizontal branch stars and yielding age constraints on the population; imaging with this camera should be a powerful complement to the VVV data. The first light was achieved in 2012 October and the Blanco DECam Bulge Survey (BDBS) has been proposed.

The *Large Synoptic Survey Telescope* has an 8.4 m aperture and will undertake multiple surveys of the southern sky in the *ugriz* bands. The precise observing program for the bulge using LSST is still under consideration. The redder bands are likely to be saturated for stars brighter than the main-sequence turnoff. The author is leading a group that will produce a bulge survey using LSST. The high S/N and spatial resolution, especially for the bluer bandpasses, may enable a very precise map of the bulge age and metallicity over the entirety of the bar and inner disk.

9.2 Spectroscopic Surveys

The Sloan Digital Sky Survey 3 (SDSS3) project has as one of its main projects the *APOGEE* survey (Allende-prieto et al. 2012; Nidever et al. 2012; Majewski 2012; <http://www.sdss3.org/surveys/apogee.php>). The survey will observe $\sim 10^5$ M giants at $R = 30,000$ in the infrared H band and will obtain composition and velocities for roughly 8,000 bulge giants. It should be possible to measure Fe and 20 other elements, but is unlikely to measure the r- and s-process elements. The great power of this method is that M giants are intrinsically luminous, and the IR is affected little by extinction. However, many of the alpha element lines, like Mg, Si, and Ca, are very strong and may yield only ~ 0.2 dex precision constraints for abundance analysis, and very extensive modeling and spectrum synthesis will be required especially at the metal-rich end (see, e.g., Origlia et al. 2002). The first data were taken in 2011 with an initial public release slated for 2013. As the Apache Point Telescope is in New Mexico, USA, the northern bulge will be targeted. There is a proposal to extend the project to the Southern Hemisphere (Sloan 4).

The GAIA-ESO survey (PI G. Gilmore) will target 10^5 stars with the FLAMES and UVES spectrographs at ESO, including 10^4 Galactic bulge stars; the project will produce a public dataset and commence in 2012.

The 4 m AAT will receive a powerful new spectrograph in 2014, called *HERMES* (Barden et al. 2010). The instrument will have three dichroic-fed channels spanning 1,000 Å and will measure both α and heavy elements. The initial GALA survey avoids the bulge, but should be capable of producing powerful survey datasets for K giants, but will be unlikely to be able to work in relatively low-extinction bulge fields (the K giants are $V = 15\text{--}16$, and optical spectra are overwhelmed by molecular bands in cool, metal-rich stars).

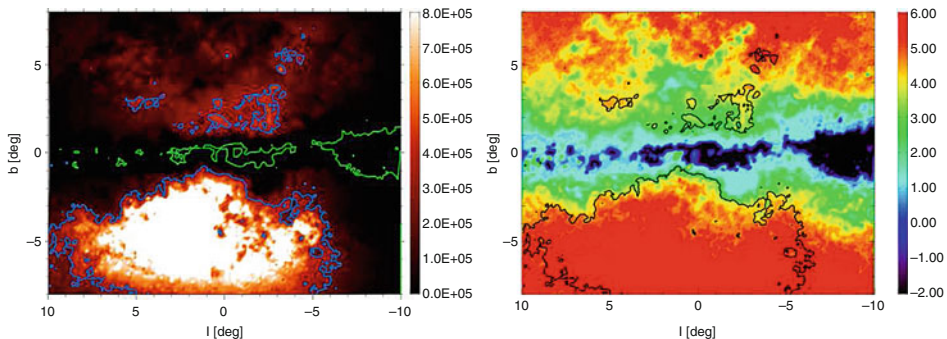
Several projects are in the planning stages that propose to employ focal planes of thousands of fibers. The 4MOST project, covering a degree-scale focal plane with $\sim 4,000$ fibers (de Jong et al. 2012), is proposed to obtain spectra at $R \sim 5,000$ and with the aim of obtaining huge samples of stars with detailed abundance analysis; this scale of survey has the potential to settle the issue of chemodynamic population complexity in the bulge. A proposed high resolution infrared spectrograph for the VLT, MOONS, would enable precision stellar abundances throughout the bulge, nearly to the Galactic Center.

9.3 Radio Surveys

With the operations of the extended VLA and ALMA (a Southern hemisphere facility), there is the possibility of undertaking significantly larger surveys of the OH/IR and SiO star population. One may refer to [Fig. 6-29d, e](#) (Habing et al. 2006) to appreciate the potential of these techniques. The Very Long Baseline Array can in principle measure parallaxes and proper motions for the SiO maser stars, but the numbers possible for reasonable observing times remain small. If sensitivities increase, this may become a very powerful technique capable of reaching precisions greater than that of GAIA.

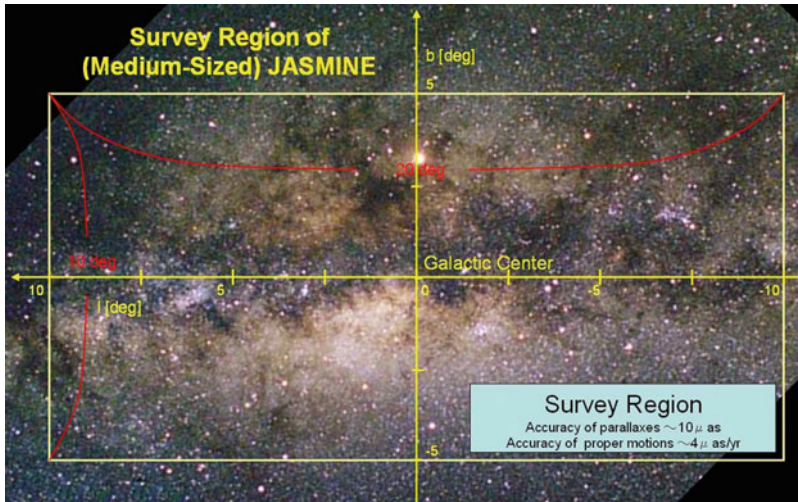
9.4 Space-Based Surveys

The GAIA survey is the main astrometric space survey project of the 2010 decade and is expected to conclude its survey by 2020, with launch in 2013. Simulations by Reyle et al. (2008; [Fig. 6-57](#)) indicate that GAIA may be able to measure 23,000,000 stars over 200 sq. deg. of the bulge. A surprising number of radial velocities will be measurable as well (to $G = 17$ mag). There is the potential to provide astrometry to $20 \mu\text{m}$ as precision in at least the outer Galactic bulge (and mission specifications discuss crowding limits of $\sim 1,000,000$ stars per sq. deg.). Much will depend on how GAIA actually performs in very crowded fields, but since the X-shaped bulge becomes most important $b < -6^\circ$ (McWilliam and Zoccali 2010; Saito et al. 2011), GAIA



■ Fig. 6-57

(Left) Density of the bulge in the GAIA astrometric fields at the limiting magnitude $G = 20$, as a function of latitude and longitude. The *blue contour* shows the iso-density of $600,000 \text{ stars deg}^{-2}$, which is the crowding limit in the astrometric fields. The *green contour* shows the iso-density of $100 \text{ stars deg}^{-2}$. (Right) Absolute magnitude M_V of bulge stars just reached at the limiting magnitude $G = 20$ (Reyle et al. 2008). GAIA may have significant impact on bulge fields at $b < -6^\circ$ where the X-shaped structure is important



■ Fig. 6-58

The JASMINE project (Gouda et al. 2002) is scoped as a series of increasingly ambitious satellites. Nano-JASMINE is a 50 cm-diameter satellite with a 5 cm mirror that will have 3 mas accuracy at $z = 7.5$ mag. The “small” JASMINE satellite will operate in the K band and achieve 10 mas/year for parallax and 4–8 mas year⁻¹ for proper motion in 3°-scale bulge fields; the aim is to launch in 2015. The main JASMINE satellite is a fully infrared mission, with an 80 cm mirror and is slated for 2022, and is to achieve 10 μ m as parallaxes over the 10° × 20° area of the bulge

stands an excellent chance of making a significant contribution to the solution of that problem. Even the northern bulge, which suffers from extinction, will be observable for GAIA because its capability to reach faint magnitudes. The red clump will still be observable for GAIA with ~ 3 –4 mag of extinction. The GAIA–ESO program and 4MOST are aimed at providing the complementary high-dispersion spectroscopy.

The JASMINE project is ambitious and may not reach its final (large satellite) form until the 2020s but offers perhaps the best opportunity in our lifetime to achieve a breakthrough in the understanding of major problems, especially the spatial distribution of stars in the X-shaped bulge, a true survey of all of the orbit families that sustain the shape of the bulge, the possibility of a separate nuclear component, and the interplay between bulge and disk in the inner galaxy. Only observations in the plane, with proper motions and parallaxes, can address the population separation problem between the disk and the bulge. The parallax will permit a direct measurement of the shape of the bulge without resorting to model fitting. ▶ Figure 6-58 shows the ambitious field of regard contemplated for the final Jasmine mission.

In summary, GAIA and JASMINE are missions that are critical to the understanding of the Galactic bulge and are likely to provide major breakthroughs.

10 Observational Challenges for the Future

The structure of the bar is presently constrained from modeling of gas dynamics, stellar surface brightness, and stellar dynamics. If surveys can yield proper motion and parallaxes for much of the bulge, we may yet get our 3D model including some metallicity information. One would

like to know whether the X-shape contains most of the mass of the bulge, and which orbits support it. There is also the question of whether the long bar is a distinct structure or part of the main bulge.

In principle, one wants to see a holistic approach to the inner galaxy. One cannot model the light of the bulge in isolation from the disk (Freudenreich 1998), and indeed, it is likely that the bulge formed from the buckling of a preexisting massive disk. The proto-Milky Way might have resembled M82 to some large extent, and to fully appreciate the results of such a history, our surveys must push well beyond the boundaries of the bulge; Ness and Freeman (2012) have shown the way, in this respect. We will need thousands of stars observed in the inner 50 pc as well.

Composition surveys of bar stars are still in their infancy. Large datasets are being built for GAIA/ESO survey and APOGEE. The host of arguments regarding subpopulations, and the detailed composition of the bulge compared with the thin and thick disks, should be settled if samples exceeding 10,000 stars, including kinematics and abundances, are developed. Projects of the scale of 4MOST could produce 100,000 bulge giants with high-resolution composition measurements.

Correlating chemical tagging with kinematics could also be interesting in terms of constraining the formation scenario, as well as the orbit classes responsible for the shape of the bar.

The tension between a shallow mass function for the bulge (star count-based) versus apparently steep mass functions for giant ellipticals (e.g., van Dokkum and Conroy 2010) needs resolution. Is the bulge different from these massive spheroids in yet another important respect, or will deeper observations of the bulge find greater numbers of low mass stars?

In a white paper of the United States Decadal Survey, Rich et al. (2009) raised a number of questions as challenges for future research on the bulge. These problems remain current; the author has added a few additional ones:

1. Did the bulge form from a single enrichment event, or did it merge from chemically distinct subcomponents?
2. Are the microlensed dwarfs from the same population as the giants?
3. Why is the bulge composition not internally consistent with predictions from massive star nucleosynthesis? Did massive rotating stars play a critical role in early chemical evolution?
4. Was the chemical evolution of the bulge brief (<1 Gyr) or more extended?
5. How can the outer bulge exhibit an abundance gradient if the evidence supports formation via a dynamical process alone?
6. Has the bulge accreted small stellar systems. If so, where is the evidence?
7. How does the composition of the bulge relate to other Galactic populations, especially the inner disk and Galactic center?
8. Does the bulge exhibit an age metallicity-composition relationship?
9. What is the nature/origin/chemistry of subpopulations in the bulge, including the X-structure?
10. Were winds or other physical factors most important in the early chemical evolution of the bulge?
11. Is the faint mass function of the bulge shallower than for giant elliptical galaxies?

Within 5 years of the publication of this chapter, the field will experience dramatic changes as the sample sizes of stars with high-precision spectroscopy and kinematics increase by an order of magnitude, and the VVV data goes public. The study of the bulge will undergo a greater transformation than at any other time in the last century.

The author is grateful for comments from Ortwin Gerhard, Christian Johnson, Will Clarkson, Livia Origlia, Juntao Shen, and Andreas Koch and for assistance from Christine Black on some aspects of the manuscript. The author acknowledges support from the US National Science Foundation grants AST-0709479 and 1212095. This publication makes use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. This research has also made use of the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. This chapter is dedicated to the memories of Albert Whitford and Victor Blanco, whose efforts at Cerro Tololo helped to spawn the modern era of bulge studies.

Cross-References

- ◆ [Astrophysics of Galactic Charged Cosmic Rays](#)
- ◆ [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- ◆ [Dynamics of Disks and Warps](#)
- ◆ [Galactic Distance Scales](#)
- ◆ [Galactic Neutral Hydrogen](#)
- ◆ [Globular Cluster Dynamical Evolution](#)
- ◆ [High-Velocity Clouds](#)
- ◆ [History of Dark Matter in Galaxies](#)
- ◆ [Interstellar PAHs and Dust](#)
- ◆ [Magnetic Fields in Galaxies](#)
- ◆ [Mass Distribution and Rotation Curve in the Galaxy](#)
- ◆ [Star Counts and the Nature of the Galactic Thick Disk](#)
- ◆ [The Infrared Galaxy](#)

References

- Abadi, M. G., Navarro, J. F., Steinmetz, M., & Eke, V. R. 2003, *ApJ*, 597, 21
- Alard, C. 2001, *A&A*, 377, 389
- Alcock, C., Allsman, R. A., Alves, D., et al. 1997, *ApJ*, 479, 119
- Alves-Brito, A., Melendez, J., Asplund, M., Ramirez, I., & Yong, D. 2010, *A&A*, 513, A35
- Arendt, R. G., Berriman, G. B., Boggess, N., et al. 1994, *ApJ*, 425, L85
- Arp, H. C. 1965, *ApJ*, 141, 43
- Athanassoula, E. 2005, *MNRAS* 358, 1477
- Athanassoula, E. 2008, *Formation and Evolution of Galaxy Disks ASP Conference Series*, Vol. 396, Proceedings of the conference held 1–5 October, 2007 at the Centro Convegno Matteo Ricci, Rome, Italy. Edited by José G. Funes, S.J., and Enrico Maria Corsini. San Francisco: Astronomical Society of the Pacific, 333
- Athanassoula, E. 2012, in *EPJ Web of Conf.* 19, *Assembling the Puzzle of the Milky Way*, ed. Le Grand-Bornand, France, ed. C. Reyle, A. Robin, & M. Schultheis, id.06004, 19, 6004
- Azzopardi, M., Rebeiro, E., Lequeux, J., & Westerglund, B. E. 1991, *AAPS*, 88, 265
- Baade, W. 1944, *ApJ*, 100, 137
- Baade, W. 1951, *Pub. Obs. Univ. Mich.*, 10, 7
- Baade, W. 1958, *Ricerche Astronomiche, Specola Vaticana*, Proceedings of a Conference at Vatican Observatory, Vol. 5, Castel Gandolfo, ed. D. J. K. O'Connell (Amsterdam: North-Holland; New York: Interscience), 303

- Babusiaux, C., & Gilmore, G. 2005, *MNRAS*, 358, 1309
- Babusiaux, C., Gomez, A., Hill, V., et al. 2010, *A&A*, 519, A77
- Ballero, S. K., Matteucci, F., Origlia, L., Rich, R. M. 2007, *A&A*, 467, 123
- Barden, S. C., Jones, D. J., Barnes, S. I., et al. 2010, *SPIE*, 7735
- Beaton, R. L., Majewski, S. R., Guhathakurta, P., et al. 2007, *ApJ*, 658, L91
- Beaulieu, S. F., Freeman, K. C., Kalnajs, A. J., Saha, P., & Zhao, H. 2000, *AJ*, 120, 855
- Becklin, E. E., & Neugebauer, G. 1968, *ApJ*, 151, 145
- Bekki, K., & Tsujimoto, T. 2011, *MNRAS*, 416, L60
- Benjamin, R. A., Churchwell, E., Babler, B. L., et al. 2005, *ApJ*, 630, L149
- Bensby, T., Feltzing, S., Ljunstrom, I., & Ilyin, I. 2005, *A&A*, 433, 185
- Bensby, T., Adén, D., Meléndez, J., et al. 2011, *A&A*, 533, A134
- Bensby, T., Yee, J. C., Feltzing, S., Johnson, J. A. et al. 2012, arXiv:1211.6848; *A&A* in press.
- Binney, J., Gerhard, O. E., Stark, A. A., Bally, J., & Uchida, K. I. 1991, *MNRAS*, 252, 210
- Binney, J., Gerhard, O., & Spergel, D. 1997, *MNRAS*, 288, 365
- Binney, J. 2009, in *IAU Symp. 254, The Galaxy Disk in Cosmological Context, Proceedings of the International Astronomical Union, Copenhagen*, ed. J. Andersen, J. Bland-Hawthorn, & B. Nordström (Cambridge/New York: Cambridge University Press), 145
- Bissantz, N., & Gerhard, O. 2002, *MNRAS*, 330, 591
- Blanco, V. M., & Blanco, B. M. 1997, *AJ*, 114, 2596
- Blanco, B. M. 1984, 89, 1836
- Blanco, V. M., McCarthy, M. F., & Blanco, B. M. 1984, *AJ*, 89, 636
- Blitz, L., & Spergel, D. N. 1991, *ApJ*, 379, 631
- Blommaert, J. A. D. L., & Groenewegen, M. A. T. 2007, in *From Stars to Galaxies: Building the Pieces to Build Up the Universe*, Vol. 374, ed. A. Vallenari et al. (San Francisco: ASP), 193
- Blum, R. D., Carr, J. S., Depoy, D. L., Sellgren, K., & Terndrup, D. M. 1994, *ApJ*, 422, 111
- Blum, R. D. 1995, *ApJ*, 444, L89
- Branch, D., Bonnell, J., & Tomkin, J. 1978, *ApJ*, 225, 902
- Brown, T. M., Ferguson, H. C., Smith, E., et al. 2003, *ApJ*, 592, L17
- Brown, T. M., Smith, E., Ferguson, H. C., et al. 2006, *ApJ*, 652, 323
- Brown, T. M., Sahu, K., Anderson, J., et al. 2010, *ApJ*, 725, L19
- Bureau, M., & Freeman, K. C. 1999, *AJ*, 118, 126
- Burkert, A., & Smith, G. H. 1997, *ApJ*, 474, L15
- Burton, W. B., & Liszt, H. S. 1978, *ApJ*, 225, 815
- Burton, W. B., Liszt, H. S. 1993, *A&A*, 274, 765
- Cabrera-Lavers, A., Hammersley, P. L., Gonzalez-Fernandez, C., et al. 2007, *A&A*, 465, 825
- Castro, S., Rich, R. M., McWilliam, A., et al. 1996, *AJ*, 111, 2439
- Castro, S., Rich, R. M., Grenon, M., Barbuy, B., & McCarthy, J. K. 1997, *AJ*, 114, 376
- Catchpole, R. M., Whitelock, P. A., & Glass, I. S. 1990, *MNRAS*, 247, 479
- Cescutti, G., & Matteucci, F. 2011, *A&A*, 525, 126
- Cescutti, G., Matteucci, F., McWilliam, A., & Chiappini, C. 2009, *A&A*, 505, 605
- Chiappini, C., Frischknecht, U., Meynet, G., et al. 2011, *Nature*, 472, 454
- Clarkson, W., et al. 2008, *ApJ*, 684, 1110
- Clarkson, W. I., Sahu, K. C., Anderson, J., et al. 2011, *ApJ*, 735, 37
- Collinge, M., Sumi, T., & Fabrycky, D. 2006, *ApJ*, 651, 197
- Combes, F. 2009, *ASP Conf Ser*, 419, 231
- Combes, F., Debbasch, F., Friedli, D., & Pfenninger, D. 1990, *A&A*, 233, 82
- Coté, P. 1999, *AJ*, 118, 406
- Cunha, K., Smith, V. V., & Gibson, B. K. 2008, *ApJ*, 679, L17
- Dame, T. M. 1999, in *The Physics and Chemistry of the Interstellar Medium*, ed. V. Ossenkopf, J. Stutzki, & G. Winnewisser (Herdecke: GCA), 100
- Dame, T. M., Hartmann, D., & Thaddeus, P. 2001, *ApJ*, 547, 792
- Debattista, V. P., Gerhard, O., & Sevenster, M. N. 2002, *MNRAS*, 334, 355
- Dehnen, W., 2000, *AJ*, 119, 800
- Debattista, V. P., and Shen, J. 2007, *ApJ*, 654, L127
- de Jong, R. S., Chiappini, C., & Schnurr, O. 2012, in *EPJ Web of Conf. 19, Assembling the Puzzle of the Milky Way*, Le Grand-Bornand, France, ed. C. Reyle, A. Robin, & M. Schultheis, id.09004, 19, 9004
- Dekel, A., & Woo, J. 2003, *MNRAS*, 344, 1131
- de Propriis, R., et al. 2011, *ApJ*, 732, L63
- de Vaucouleurs, G. 1964, *The Galaxy and the Magellanic Clouds*, Vol. 20, ed. F. J. Kerr, & A. W. Rodgers (Canberra: Australian Academy of Science), 195
- de Vaucouleurs, G., & Pence, W. D. 1978, *AJ*, 83, 1163
- Dwek, E., Arendt, R. G., Hauser, M. G., et al. 1995, *ApJ*, 445, 716
- Eggen, O. J., Lynden-Bell, D., & Sandage, A. R. 1962, *ApJ*, 136, 748
- Eisenhauer, F., Genzel, R., Alexander, T., et al. 2005, *ApJ*, 628, 246
- Eisenhauer, F., Schödel, R., Genzel, R., et al. 2003, *ApJ*, 597, L121
- Elmegreen, B. G. 1999, *ApJ*, 517, 103

- Englmaier, P., & Gerhard, O. 1999, *MNRAS*, 304, 512
- Erwin, P., and Sparke, L. S. 2002, *AJ*, 124, 65
- Feast, M. 1963, *MNRAS*, 125, 367
- Feast, M. W. 1997, *MNRAS*, 284, 761
- Feltzing, S., & Gilmore, G. 2000, *A&A*, 355, 949
- Fernley, J. A., Jameson, R. F., Longmore, A. J., Watson, F. G., & Wesselink, T. 1987, *MNRAS*, 226, 927
- Ferraro, F. R., et al. 2009, *Nature*, 462, 483
- Fich, M., Blitz, L., & Stark, A. 1989, *ApJ*, 342, 272
- Figer, D. F., Rich, R. M., Kim, S. S., Morris, M., & Serabyn, E. 2004, *ApJ*, 601, 319
- Freudenreich, H. T., 1998, *ApJ*, 492, 495
- Frogel, J. A. 1988, *ARA&A*, 26, 51
- Frogel, J. A., & Elias, J. H. 1988, *ApJ*, 324, 823
- Frogel, J. A., & Whitford, A. E. 1987, *ApJ*, 315, 199
- Frogel, J. A., & Whitelock, P. A. 1998 *AJ*, 116, 754
- Fukugita, M., Hogan, C. J., & Peebles, P. J. E. 1998, *ApJ*, 503, 518
- Fulbright, J. P., McWilliam, A., & Rich, R. M. 2006, *ApJ*, 636, 821
- Fulbright, J. P., McWilliam, A., & Rich, R. M. 2007, *ApJ*, 661, 1152
- Fux, R. 1999, *A&A*, 345, 787
- Gallino, R., Arlandini, C., Busso, M., et al. 1998, *ApJ*, 497, 388
- Genzel, R., Pichon, C. Eckart, A., Gerhard, O. E., & Ott, T. 2000, *MNRAS*, 317, 348
- Gerhard, O. 2002, *The Dynamics, Structure and History of Galaxies: A Workshop in Honour of Professor Ken Freeman*, ASP Conf. Ser. 273, ed. G. S. Da Costa and Helmut Jerjen. ISBN: 1-58381-114-1. (San Francisco: ASP), 73
- Gerhard, O. 2011, *Memorie della Societa Astronomica Italiana Supplementi*, 18, 185
- Gerhard, O., & Martinez-Valpuesta, I. 2012, *ApJ*, 744, L8
- Ghez, A. M., Duchene, G., Matthews, K., et al. 2003, *ApJ*, 586, L127
- Ghez, A. M., Salim, S., Weinberg, N. N., et al. 2008, *ApJ*, 689, 1044
- Gillessen, S., Eisenhauer, F., Trippe, S., et al. 2009, *ApJ*, 692, 1075
- Gonzalez, O. A., Rejkuba, M., Zoccali, M., et al. 2011a, *A&A*, 530, A54
- Gonzalez, O. A., Rejkuba, M., Zoccali, M., Valenti, E., & Minniti, D. 2011b, *A&A*, 534, A3
- Gonzalez, O. A., Rejkuba, M., Minniti, D., et al. 2011c, *A&A*, 534, L14
- Gonzalez, O. A., Rejkuba, M., Zoccali, M., et al. 2012, *A&A*, 543, A13
- Gouda, N., Tsujimoto, T., Kobayashi, Y., et al. 2002, *Astrophys Space Sci*, 280, 89
- Gould, A., Bahcall, J. N., & Flynn, C. 1997, *ApJ*, 482, 913
- Graham, A. W., Spitler, L. R., Forbes, D. A., et al. 2012, arXiv:1203.3608
- Greggio, L., & Renzini, A. 1990, *ApJ*, 364, 35
- Groenewegen, M. A. T., & Blommaert, J. A. D. L. 2005, *A&A*, 443, 143
- Guarnieri, M. D., Renzini, A., & Ortolani, S. 1997, *ApJ*, 477, L21
- Habing, H. J., Sevenster, M. N., Messineo, M., van de Ven, G., & Kuijken, K. 2006, *A&A*, 458, 151
- Hamadache, C., Le Guillou, L., Tisserand, P., et al. 2006, *A&A*, 454, 185
- Hammersley, P. L., Lopez-Corredoira, M., & Garzon, F. 2001, in *Tetons 4: Galactic Structure, Stars and the Interstellar Medium*, Vol. 231 ed. C. E Woodward, M. D. Bica, & J. M. Shull (San Francisco, CA: ASP), 81
- Hartmann, D., & Burton, W. B. 1997, *Atlas of Galactic Neutral Hydrogen*, ed. Dap Hartmann, W. Butler Burton (Cambridge, UK: Cambridge University Press), 243, ISBN 0521471117
- Hartwick, F. D. A. 1976, *ApJ*, 209, 418
- Hill, V., Lecureur, A., et al. 2011, *A&A*, 534, 80
- Holtzman, J. A., Light, R. M., Baum, W. A., et al. 1993, *AJ*, 106, 1826
- Holtzman, J. A., Watson, A. M., Baum, W. A., et al. 1998, *AJ*, 115, 1946
- Howard, C. D., Rich, R. M., Reitzel, D. B., Koch, A., de Propriis, R., & Zhao, H. 2008, *ApJ*, 688, 1060
- Howard, C. D., Rich, R. M., Clarkson, W. I., et al. 2009, *ApJ*, 702, 153
- IAU 153, 1993. *Galactic Bulges: Proceedings of the 153rd Symposium of the International Astronomical Union held in Ghent, Belgium August 17–22*. H. Habing and H. Dejonghe, eds. Kluwer Academic
- IAU 245, 2008. *Formation and Evolution of Galactic Bulges*, IAU Symposium 245, M. Bureau, E. Athanassoula, B. Barbuy eds. Cambridge U. press.
- Ibata, R. A., & Gilmore, G. 1995, *MNRAS*, 275, 591
- Ibata, R. A., Gilmore, G., & Irwin, M. J. 1994, *Nature*, 370, 194
- Immeli, A., Samland, M., Gerhard, O., & Westera, P. 2004, *A&A*, 413, 547
- Johnson, C. I., & Pilachowski, C. 2010, *ApJ*, 722, 1373
- Johnson, C. I., Rich, R. M., Fulbright, J. P., Valenti, E., & McWilliam, A. 2011, *ApJ*, 732, 108
- Johnson, C. I., Rich, R. M., Kobayashi, C., & Fulbright, J. P. 2012a, *ApJ*, 749, 175
- Johnson, C. I., Rich, R., Kunder, A., et al. 2012b, *American Astronomical Society Meeting Abstracts*, 219, #152.11
- Johnson, C. I., Rich, R. M., et al. 2012c (in preparation)

- Kent, S. M. 1989, *AJ*, 97, 1614
- Kent, S. M. 1992, *ApJ*, 387, 181
- Kent, S. M., Dame, T. M., & Fazio, G. 1991, *ApJ*, 378, 131
- Kerr, F. J., Bowers, P. F., Jackson, P. D., & Kerr, M. 1986, *A&AS*, 66, 373
- Kormendy, J., & Illingworth, G. 1982, *ApJ*, 256, 460
- Kormendy, J. & Kennicutt, R. C., Jr. 2004, *ARA&A*, 42, 603
- Kormendy, J., Drory, N., Bender, R., & Cornell, M. E. 2010, *ApJ*, 723, 54
- Kuijken, K., & Merrifield, M. R. 1995, *ApJ*, 443, L13
- Kuijken, K., & Rich, R. M. 2002, *AJ*, 124, 2054
- Kunder, A. & Chaboyer, B. 2008, *AJ*, 136, 2441
- Kunder, A., Koch, A., Rich, R. M., et al. 2012, *AJ*, 143, 57
- Lanzoni, B. et al. *ApJ*, 2010, 717, 653
- Lauer, T., Bender, R., Kormendy, J., Rosenfield, P., & Green, R. F. 2012, *ApJ*, 745, 121
- Launhardt, R., Zylka, R., & Mezger, P. G. 2002, *A&A*, 384, 112
- Lecureur, A., Hill, V., Zoccali, M. et al. 2007, *A&A*, 465, 799
- Li, Z. Y., and Shen, J. 2012, *ApJ*, 757, L7
- Lindqvist, M., Habing, H. J., & Winnberg, A. 1992a, *A&A*, 259, 118
- Lindqvist, M., Winnberg, A., Habing, H. J., & Matthews, H. E. 1992b, *A&A Suppl.* 92, 43
- Liszt, H. S., & Burton, W. B. 1980, *ApJ*, 236, 779
- Lloyd Evans, T. 1976, *MNRAS*, 174, 169
- López-Corredoira, M., Cabrera-Lavers, A., & Gerhard, O. E. 2005, *A&A*, 439, 107
- López-Corredoira, M., Cabrera-Lavers, A., Mahoney, T. J., et al. 2007, *ApJ*, 133, 154
- Maeder, A. 1992, *A&A*, 264, 105
- Maihara, T., Oda, N., Sugiyama, T., & Okuda, H. 1978, *PASJ*, 30, 1
- Majewski, S. 2012, The Chemical Evolution of the Milky Way, held January 23–27, 2012 at the Sexten Center for Astrophysics (SCfA), Sesto Pusteria, Dolomites, Bolzano, Italy. Online at http://www.sexten-cfa.eu/it/conferenze/conferences2012/details/17-Chem_evo_milky_way, id.24
- Maraston, C., Greggio, L., Renzini, A., et al. 2003, *A&A*, 400, 823
- Marin-Franch, A., Cassisi, S., Aparicio, A., & Pietrinferni, A. 2010, *APJ*, 714, 1072
- Martinez-Valpuesta, I., & Gerhard, O., 2011, *ApJ*, 734, L20
- Matsumoto, T., et al. 1982 in *The Galactic Center*, ed. G. Riegler & R. Blandford (New York: American Institute of Physics), 48
- Matteucci, F., & Brocato, E. 1990, *ApJ*, 365, 539
- Matteucci, F., Romano, D., & Molaro, P. 1999, *A&A*, 341, 458
- McClean, I., Becklin, E. E., Bendiksen, O., et al. 1998, *SPIE*, 3354, 566
- McWilliam, A. 1997, *ARAA*, 35, 503
- McWilliam, A., & Rich, R. M. 1994, *APJS*, 91, 749 (MR94)
- McWilliam, A., & Rich, R. M. 2004, *Origin and Evolution of the Elements* (Cambridge, UK/New York: Cambridge University Press), 38
- McWilliam, A. & Zoccali, M. 2010, *ApJ*, 724, 1491
- McWilliam, A., Rich, R. M., & Smecker-Hane, T. A. 2003, *ApJ*, 592, L21
- McWilliam, A., Matteucci, F. M., Ballero, S., et al. 2008, *AJ*, 136, 367
- McWilliam, A., Fulbright, J. P., & Rich, R. M. 2010, *IAUS*, 265, 279
- Melendez, J., Asplund, M., Alves-Brito, A., et al. 2008, *A&A*, 484, L21
- Menéndez-Delmestre, K., Sheth, K., Schinnerer, E., Jarrett, T. H., & Scoville, N. Z. 2007, *ApJ*, 657, 790
- Merrifield, M. R. 2004, in *Milky Way Surveys: The Structure and Evolution of our Galaxy*, ed. D. Clemens, R. Shah, & T. Brainerd, *ASP Conf. Ser.*, 317, 289
- Messineo, M. 2002, *AJ*, *A&A* 393, 115
- Minniti, D. 1995a, *AJ*, 109, 1663
- Minniti, D. 1995b, *A&A*, 300, 109
- Minniti, D. 1996a, *ApJ*, 459, 175
- Minniti, D. 1996b, *ApJ*, 459, 579
- Minniti, D. 1996c, *ApJ*, 459, 175
- Minniti, D., White, S. D. M., Olszewski, E. W., & Hill, J. M. 1992, *ApJ*, 393, L47
- Minniti, D., Olszewski, E. W., Liebert, J., et al. 1995, *MNRAS*, 277, 1293
- Minniti, D., Liebert, J., Olszewski, E. W., & White, S. D. M. 1996, *AJ*, 112, 590
- Minniti, D., Vandehei, T., Cook, K. H., Griest, K., & Alcock, C. 1998, *ApJ*, 499, L175
- Minniti, D., Lucas, P. W., Emerson, J. P., et al. 2010, *New Astron.*, 15, 433
- Morris, M. & Serabyn, E. 1996, *ARA&A*, 34, 645
- Mouhcine, M., Ibata, R., & Rejkuba, M. 2010, *ApJ*, 714, L12
- Mould, J. R. 1983, *ApJ*, 266, 255
- Mould, J. R. 1984, *PASP*, 96, 773
- Nakada, Y., Onaka, T., Yamamura, I., et al. 1991, *Nature*, 353, 140
- Nataf, D. M., & Gould, A. P. 2012, *APJL*, 751, L39
- Nassau, J. J. & Blanco, V. M. 1958, *ApJ*, 128, 46
- Nataf, D. M., Udalski, A., Gould, A., Fouqué, P., & Stanek, K. Z. 2010, *ApJ*, 721, L28
- Nataf, D. M., Udalski, A., Gould, A., & Pinsonneault, M. H. 2011, *ApJ*, 730, 118
- Nataf, D. M., Gould, A., & Pinsonneault, M. H. 2012, *astro-ph/1203.5791N*
- Ness, M., Freeman, K., Athanassoula, E., et al. 2012, *ApJ*, 756, 22

- Ness, M., & Freeman, K. 2012, in EPJ Web of Conf., Vol. 19, Assembling the Puzzle of the Milky Way, Le Grand-Bornand, France, ed. C. Reyle, A. Robin, & M. Schultheis, id.06003, 19, 6003
- Nidever, D. L., Zasowski, G., Majewski, S. R., et al. 2012, *ApJ*, 755, L25
- Nishiyama, S., Nagata, T., Sato, S., et al. 2006, *ApJ*, 647, 1093
- Norman, C. A., Sellwood, J., & Hassan, H. 1996, *ApJ*, 462, 114
- O'Connell, R. W. 1999, *ARA&A*, 37, 603
- Olsen, K. A. G., Blum, R. D., Stephens, A. W., et al. 2006, *AJ*, 132, 271
- Origlia, L., Rich, R. M., & Castro, S. 2002, *AJ*, 123, 1559
- Origlia, L., Rich, R. M., Ferraro, F., et al. 2011, *ApJ*, 726, L20
- Ortolani, S., Renzini, A., et al. 1995, *Nature*, 377, 701
- Ortolani, S., Barbuy, B., Momany, Y., et al. 2011, *ApJ*, 737, 31
- Paczynski, B. 1986, *ApJ*, 304, 1
- Paczynski, B., & Stanek, K. Z. 1998, *ApJ*, 494, L219
- Paczynski, B., Stanek, K. Z., Udalski, A., et al. 1994, *ApJ*, 435, L113
- Peterson, R. 1976, *ApJS*, 30, 61
- Pompèia, L., Barbuy, B., & Grenon, M. 2003, *ApJ*, 592, 1173
- Proctor, R. N., Sansom, A. E., & Reid, I. N. 2000, *MNRAS*, 311, 37
- Puzia, T. H., Saglia, R. P., Kissler-Patig, M., et al. 2002, *A&A*, 395, 45
- Raha, N., Sellwood, J. A., James, R. A., & Kahn, F. D. 1991, *Nature*, 352, 411
- Ramirez, S. V., Stephens, A. W., Frogel, J. A., & DePoy, D. L. 2000, *AJ*, 120, 833
- Rangwala, N., Williams, T. B., & Stanek, K. Z. 2009, *ApJ*, 691, 1387
- Rattenbury, N. J., Mao, S., Debattista, V. P., et al. 2007a, *MNRAS*, 378, 1165
- Rattenbury, N. J., Mao, S., Sumi, T., & Smith, M. C. 2007b, *MNRAS*, 378, 1064
- Reid, I. N., & Gizis, J. E. 1997, *AJ*, 113, 2246
- Reid, M. J., Schneps, M. H., Moran, J. M., et al. 1988, *ApJ*, 330, 809
- Renzini, A. 1994, *A&A*, 285, L5
- Renzini, A., & Fusi Pecci, F. 1988, *ARA&A*, 26, 199
- Renzini, A. 1999, in *Cambridge Contemporary Astrophysics, The Formation of Galactic Bulges*, ed. C. M. Carollo, H. C. Ferguson, R. F. G. Wyse (Cambridge, UK/New York: Cambridge University Press), 9
- Reyle, C., Marshall, D. J., Schultheis, M., & Robin, A. C. 2008, *SF2A-2008*, 29
- Rich, R. M. 1988, *AJ*, 95, 828
- Rich, R. M., Mould, J., Picard, A., Frogel, J. A., & Davies, R. 1989, *ApJ*, 341, L51
- Rich, R. M. 1990, *ApJ*, 362, 604
- Rich, R. M., & McWilliam, A. 2000, *SPIE*, 4005, 150 (RM00)
- Rich, R. M., & Mould, J. R. 1991, *AJ*, 101, 1286
- Rich, R. M., & Origlia, L. 2005, *ApJ*, 634, 1293
- Rich, R. M., Reitzel, D. B., Howard, C. D., & Zhao, H. 2007, *ApJ*, 658, L29
- Rich, R. M., Origlia, L., & Valenti, E. 2007b, *ApJ Lett*, 665, L119
- Rich, R. M. 2008 in *The Metal Rich Universe*, ed. G. Israelian and G. Meynet (Cambridge: Cambridge University Press), 3
- Rich, R. M., Howard, C., Reitzel, D. B., Zhao, H., & de Propris, R. 2008, *IAU Symp.*, 245, 333
- Rich, R. M., Clarkson, W., Cohen, J., Howard, C., McWilliam, A., Johnson, J., Johnson, C., Cunha, K., Smith, V., Fulbright, J. 2009, *Astro2010: The Astronomy and Astrophysics Decadal Survey*, Science White Papers, no. 246
- Rich, R. M., R.M., Origlia, L., & Valentia, E. 2012, *ApJ*, 746, 59
- Rich, R. M., et al. 2013 (in preparation)
- Robin, A. C., Reylé, C., Derrière, S., & Picaud, S. 2003, *A&A*, 409, 523
- Robin, A. C., Marshall, D. J., Schultheis, M., & Reyle, C. 2012, *A&A*, 538, A106
- Rodriguez-Fernandez, N. J., & Combes, F. 2008, *A&A*, 489, 115
- Ryde, N., Edvardsson, B., Gustafsson, B. 2009, *A&A*, 496, 701
- Ryde, N., Gustafsson, B., Edvardsson, B., et al. 2010, *A&A*, 509, A20
- Sadler, E. M., Rich, R. M., & Terndrup, D. M. 1996, *AJ*, 112, 171
- Saglia, R. P., Fabricius, M., Bender, R., et al. 2010, *A&A*, 509, A61
- Saha, K., Martinez-Valpuesta, I., & Gerhard, O. 2012, *MNRAS*, 421, 333
- Sahu, K. C., Casertano, S., Bond, H. E., et al. 2006, *Nature*, 443, 534
- Saito, R. K., Zoccali, M., McWilliam, A., et al. 2011, *AJ*, 142, 76
- Saito, R. K., Hempel, M., Minniti, D., et al. 2012, *A&A*, 537, A107
- Sarajedini, A., & Jablonka, P. 2005, *AJ*, 130, 1627
- Searle, L., & Sargent, W. 1972, *ApJ*, 173, 25
- Sellwood, J. A., & Wilkinson, A. 1993, *Rep Prog Phys*, 56, 173
- Sevenster, M. N., Chapman, J. M., Habing, H. J., Killeen, N. E. B., & Lindqvist, M. 1997, *A&AS*, 122, 79
- Sevenster, M., Saha, P., Valls-Gabaud, D., & Fux, R. 1999, *MNRAS*, 307, 584
- Shapley, H. 1919, *ApJ*, 49, 311
- Sharples, R., Walker, A., & Cropper, M. 1990, *MNRAS*, 246, 54

- Shectman, S. A. 1984, *SPIE*, 445, 128
- Shen, J. and Debattista, V. P. 2009, *ApJ*, 690, 758
- Shen, J. Rich, R. M., Kormendy, J., Howard, C. D., de Propriis, R., & Kunder, A. 2010, *ApJ*, 720, L72
- Shen, J. et al. 2012, arXiv 207.2872
- Sheth, K., Elmegreen, D. M., Elmegreen, B. G., et al. 2008, *ApJ*, 675, 1141
- Sjouwerman, L. O., Habing, H. J., Lindqvist, M., van Langevelde, H. J., & Winnberg, A. 1999, *The Central Parsecs of the Galaxy*, Vol. 186 (San Francisco, CA: ASP), 379
- Sjouwerman, L.O. et al. 2000, in *Star formation from the small to the large scale. ESLAB symp. (33: 1999: Noordwijk, The Netherlands)*, Proceedings of the 33rd ESLAB symposium on star formation from the small to the large scale, ESTEC, Noordwijk, The Netherlands, 2–5 November 1999 Noordwijk, ESA SP 445, ed. F. Favata, A. Kaas, & A. Wilson (The Netherlands: European Space Agency (ESA)), 519
- Soto, M., Kuijken, K., & Rich, R. M. 2007, *ApJ*, 540, 48
- Soto, M., Kuijken, K., & Rich, R. M. 2012, *A&A*, 540, A48
- Spaenhauer, A., Jones, B. F., & Whitford, A. E. 1992, *AJ*, 103, 297
- Spinrad, H., & Taylor, B. J. 1969, *ApJ*, 157, 1279
- Spinrad, H., Taylor, B.J., & van den Bergh, S. 1969, *AJ*, 74, 525
- Stanek, K. Z., Mateo, M., Udalski, A., et al. 1994, *ApJ*, 429, L73
- Stanek, K. Z., Udalski, A., Szymanski, M., et al. 1997, *ApJ*, 477, 163
- Stebbins, J., & Whitford, A. E. 1947, *AJ*, 52, 130
- Stephens, A. W., Frogel, J. A., DePoy, D. L., et al. 2003, *AJ*, 125, 2473
- Sumi, T., Wu, X., Udalski, A., et al. 2004, *MNRAS*, 348, 1439
- Sweigart, A. V., & Gross, P. G. 1978, *ApJS*, 36, 405
- Tal, T., van Dokkum, P. G., Nelan, J., & Bezanson, R. 2009, *AJ*, 138, 1417
- Terndrup, D. 1988, *AJ*, 96, 884
- Terndrup, D., Rich, R. M., & Whitford, A. E. 1984, *PASP*, 96, 796
- Terndrup, D. M., Sadler, E. M., & Rich, R. M. 1995, *AJ*, 110, 1774
- Terndrup, D. M., An, Deokkeun, Hansen, A., et al. 2004, *Ap&SS*, 291, 247
- Tinsley, B. M., & Gunn, J. E. 1976, *ApJ*, 206, 525
- Tomkin, J., & Lambert, D.L. 1983, *ApJ*, 273, 722
- Tremaine, S., Gebhardt, K., Bender, R., et al. 2002, *ApJ*, 574, 740
- Tyson, N. D., & Rich, R. M. 1991, *ApJ*, 367, 547
- Udalski, A. 1998, *Acta Astron.*, 48, 113
- Utenthaler, S., Hron, J., Lebzelter, T., et al. 2008, *A&A*, 478, 527
- van den Bergh, S., & Herbst, E. 1974, *AJ*, 79, 603
- van Dokkum, P.G., & Conroy, C. 2010, *Nature*, 468, 940
- Vanhollebeke, E., Groenewegen, M. A. T., & Girardi, L. 2009, *A&A*, 498, 95
- van Loon, J. T., Gilmore, G. F., Omont, A., et al. 2003, *MNRAS*, 338, 857
- Vieira, K., Casetti-Dinescu, D. I., Mendez, R. A., et al. 2007, *AJ*, 134, 1432
- Walker, A. R., & Terndrup, D. M. 1991, *ApJ*, 378, 119
- Wang, Y., Zhao, H. et al. 2012, *MNRAS* in press.
- Weiland, J. L., Arendt, R. G., et al. 1994, *ApJ*, 425, L81
- Weinberg, M. D. 1992, *ApJ*, 384, 81
- Wheeler, J., Sneden, C., & Truran, J. W. 1989, *ARA&A*, 27, 279
- Whitlock, P. A., & Catchpole, R. 1992, in *IAU Symp. 149, The Stellar Populations of Galaxies*, ed. B. Barbuy, & A. Renzini (Dordrecht: Kluwer), 503
- Whitlock, P., Feast, M., & Catchpole, R. 1991, *MNRAS*, 248, 276
- Whitlock, P. 1992, Long-period variables and carbon stars in the Galactic Bulge, in *Galactic bulges: Proceedings of the 153rd Symposium of the International Astronomical Union held in Ghent, Belgium, August 17–22, 1992*, International Astronomical Union. Symposium no. 153, ed. by H. DeJonghe & H. J. Habing (Dordrecht: Kluwer Academic Publishers), 39
- Whitford, A. E. 1978, *ApJ*, 226, 777
- Whitford, A. E., & Rich, R. M. 1983, *ApJ*, 274, 723
- Woosley, S. E., & Weaver, T. A. 1995, *ApJS*, 101, 181
- Wyse, R. F. G., Gilmore, G., Franx, M. 1997, *ARA&A*, 35, 637
- Zhao, H. 1996, *MNRAS*, 283, 149
- Zhao, H., Spergel, D. N., & Rich, R. M. 1994, *AJ*, 108, 2154
- Zhao, H., Rich, R. M., & Spergel, D. N. 1996, *MNRAS*, 282, 175
- Zoccali, M., Cassisi, S., Frogel, J. A., et al. 2000, *ApJ*, 530, 418
- Zoccali, M., Renzini, A., Ortolani, S., et al. 2003, *A&A*, 399, 931
- Zoccali, M., Lecureur, A., Barbuy, B., et al. 2006, *A&A*, 457, L1
- Zoccali, M., Hill, V., Lecureur, A., et al. 2008, *A&A*, 486, 177

7 Open Clusters and Their Role in the Galaxy

Eileen D. Friel

Department of Astronomy, Indiana University, Bloomington,
Indiana, USA

1	<i>Introduction and Overview</i>	348
1.1	Surveys and Catalogs	349
2	<i>Open Clusters as Stellar Laboratories</i>	352
2.1	Color-Magnitude Diagrams	352
2.2	Structural Properties and Dynamical Evolution	356
2.2.1	Structural Properties and Masses	356
2.2.2	Cluster Dynamical Evolution	357
2.3	Cluster Mass Functions	361
2.4	Stellar Evolution and Star Clusters	362
2.4.1	Convective Overshooting	362
2.4.2	White Dwarfs and the Initial–Final Mass Function	363
2.4.3	Binary Stars and Blue Stragglers	363
2.4.4	Stellar Nucleosynthesis and Evolution	364
3	<i>Open Clusters as Galactic Tracers</i>	365
3.1	Spatial Distribution of Clusters	365
3.2	Cluster Physical Parameters	368
3.3	Spiral Arms	370
3.4	Longevity of Open Clusters	370
3.5	The Oldest Open Clusters	372
4	<i>Galactic Chemical Evolution</i>	375
4.1	Disk Abundance Gradients	376
4.2	Evolution of the Abundance Gradient with Age	379
4.3	Elemental Abundance Ratios	380
4.4	Age–Metallicity Relationship	382
4.5	Comparison to the Disk Field Populations	385
4.6	Comparison to Theoretical Models	385
5	<i>Clusters in the Context of Galaxy Formation and Evolution</i>	387
	<i>References</i>	389

Abstract: Galactic open star clusters play diverse roles as probes of astrophysical phenomena on many scales. As gravitationally bound stellar systems of from several hundred to tens of thousands of stars they are useful laboratories for the investigation of issues of stellar evolution and nucleosynthesis, stellar interactions and dynamical processes, and star formation. Since open clusters exhibit a wide range of properties and are found at all ages and almost all locations in the galactic disk, when looked at as a system, they are excellent tracers of galactic structure and evolution.

This chapter introduces the properties of open clusters in the Milky Way, discussing their structure, masses, and mass functions. Open clusters are strongly affected both by internal dynamical evolution and by encounters with external forces, such as molecular clouds and the galactic tidal field. N-body simulations provide a mechanism to explore these effects which lead to significant modification of the cluster internal structure and mass and stellar distributions, and control the cluster longevity. Open clusters also provide ideal tests for the confrontation of stellar evolutionary models with observation through their wide range of ages and sampling of stellar masses.

In the context of the Milky Way galaxy, correlations of cluster properties with location provide important constraints to our understanding of both the processes of cluster formation and their dynamical evolution. The dependence of spatial distribution on age within the open cluster system points to a complex interplay between cluster formation and survivability. Abundance gradients, both of overall metallicity and individual elemental abundance ratios, and their evolution over time, point to a complex history of chemical enrichment in the galactic disk. Finally, open clusters are discussed briefly in the context of galaxy formation, mergers, and the development of the outer galactic disk.

Keywords: Galaxy: abundances, Galaxy: disk, Galaxy: open clusters and associations, Galaxy: structure, Stars: abundances, Stars: binaries: general, Stars: C-M diagrams, Stars: kinematics and dynamics, Stars: mass function

1 Introduction and Overview

Star clusters are remarkable tools for studying a vast array of astrophysical phenomena. They serve as laboratories that challenge our understanding of stellar interiors and evolution. Their characteristics as a population elucidate issues ranging from local and global star formation to the structure and evolution of galaxies. The study of star clusters in our own galaxy has provided key information in areas ranging from the details of stellar convection or radiation transport in the cores of intermediate mass stars, to the accuracy of modeling of stellar dynamical systems, to the evidence for galactic mergers and cosmological theories of hierarchical galaxy formation.

Among the populations of star clusters, open clusters play a particularly diverse role. Their range of properties makes them especially attractive as probes of these many facets of astrophysics. Unlike the globular clusters that are thought of as having relatively well-defined properties and limited range of mass, luminosity, structural characteristics and age, open clusters span a wide range of properties.

But what makes an open cluster “open”? The name first denotes the appearance, largely by reference to the much more populous and compact globular clusters. Open clusters are less massive, less centrally concentrated, and in most cases, present a sparse and dispersed looking

aggregation above the background field of the sky. A typical open cluster has on the order of 100 or fewer visible members, which leads to its discovery and classification, although its total mass may be many times this amount. Even so, open clusters present a diverse array of appearances, ranging from the quite populous examples of the 100 Myr old M11, to the archetypical M67, to the sparse collections of only a few dozens of stars that make up NGC 3680.

In fact, the characterization of an open cluster usually rests on a collection of observed and deduced properties, including not only its appearance, but its location in the galaxy, its age, its kinematics, and its chemical composition. Like any stellar population, the assignment of the appellation “open cluster” comes after a consideration of the majority of these properties being consistent with membership. As knowledge of open clusters increases, the boundaries of these classifications often become murkier, raising the question of “transition” objects that span the domains of classical globular and open clusters, and posing interesting challenges for theories of cluster formation and evolution.

This evolution of understanding and the increasingly complex picture of cluster populations will be explored in the following sections, but at the outset it is useful to start with a simplified global picture of open clusters and why these characteristics are used to help define the population as a whole.

Open clusters have historically also been called galactic clusters, primarily because of their predominant location in the disk of the Milky Way galaxy. Again, this assignment by location is based on a distinction from the globular clusters, which were considered to populate the spherically distributed halo of the galaxy. Indeed, open clusters are found distributed closely along the galactic plane, with scaleheights consistent with the thin disk stellar populations. Found at almost all galactic radii, barring observational selection effects, this distribution makes them ideal tracers of the galactic disk, probing a large range of distances and locations in the plane of the Galaxy.

The typical open cluster is also young, perhaps a few hundred million years old. But open clusters are found at all ages in the disk. The youngest clusters are being formed now, and their study illuminates the processes of star formation. At the other extreme are open clusters many billions of years old, and whose great ages both frame our understanding of cluster evolution and guide our notions of the early galactic disk. Most importantly, the wide range of ages spanned by the open cluster population affords the opportunity to probe the entire history of the galactic disk, something not provided by other stellar tracers whose populations sample only limited age ranges.

As objects that populate the galactic disk, open clusters are also typically found to have overall chemical compositions that are close to solar, again by contrast to the globular clusters that are typically more metal-poor and show nonsolar abundance patterns. A closer inspection of open clusters shows that their chemical composition across the cluster population varies in useful and interesting ways that reveal both the underlying processes of stellar evolution and nucleosynthesis and the global patterns of chemical enrichment throughout the galaxy. As with many populations, it is the extremes of these distributions that offer special insight.

1.1 Surveys and Catalogs

Of course, the ability to study open clusters as astrophysical laboratories and to use them as tracers of the broader picture of galactic structure and evolution, rests on being able to identify them reliably. In fact, it is the relative ease of detecting and determining the properties of

star clusters that has made them useful tools. Although there are challenges in detecting clusters, there are immediate advantages, too. As ensembles of stars with the brightest members revealing the top end of a mass function, clusters can be detected to substantial distances. Determination of fundamental cluster properties, such as distance, reddening, and age, is possible to greater reliability than for individual field stars, through the analysis of color-magnitude diagrams (CMDs), color-color relationships, and comparison to theoretical isochrones. Under the assumption, born out by observations, that the cluster forms from a common natal cloud within a very short period of time, the study of cluster members provides the advantage of understanding their mass, evolutionary state, luminosity, gravity, and those fundamental parameters that enable the study of stellar physics. And because one can build up samples of dozens of stars within a cluster, basic stellar and cluster parameters, such as velocity, chemical composition, and stellar activity, etc, can be determined more precisely, if not accurately, as long as membership is well understood.

There have been and continue to be many efforts to create catalogs of galactic open clusters. The first of these to have widespread use, and the basis for many catalogs and studies that followed, was that by Lynga (1981 – Catalog of open cluster data available through CDS, Strasbourg, updated and published in 1987 as the Lund Catalog of Open Cluster Data, CDS, Strasbourg). The Lynga catalog collected published cluster data on a systematic and unbiased basis and its 1981 version contained 1,180 objects. These objects were identified and collected from a wide variety of visual surveys of photographic plate collections, with an equally wide variety of selection and classification techniques, but the catalog provided the first extensive samples of open clusters and stellar associations for study. Lynga's first analysis based on this catalog (1982) established many of the fundamental cluster properties and their dependence on galactic structure. These correlations were further refined in Janes et al. (1988) who worked from a fundamental dataset of 421 clusters whose parameters had been placed on a uniform system. These two papers established many of the diagnostic trends, such as galactic gradients in metallicity, cluster longevity, cluster size and type, that have continued to be elaborated on with more complete and extensive data sets over the years. Most of the fundamental characteristics have not changed from these initial studies, as will be seen.

The next significant step came with J.-C. Mermilliod's establishment of a web-based database for galactic open clusters (WEBDA), deriving from the Base Donnees Amas which included both derived cluster properties such as age, distance, reddening, and metallicity and measurements for individual stars in the cluster fields. The database includes photometry in most photometric systems in which the cluster stars have been observed, spectral classifications, radial and rotational velocities, astrometric data, with membership probabilities, positions and, most importantly, a complete bibliography of published data and a thorough cross-identification between different studies. The database can also be queried in a variety of ways based on stellar or cluster parameters and available data. The database can be found at <http://www.univie.ac.at/webda/webda.html>. The WEBDA has been the basis for many studies of overall cluster properties, following the early work by Lynga.

The WEBDA contains only clusters for which there are individual stellar observations and derived cluster parameters; it is not a complete listing of all identified candidate clusters. That is provided by W. Dias, who has compiled and keeps current a catalog of all identified clusters and cluster candidates published in the literature (Dias et al. 2002). Building from the Lynga and Mermilliod catalogs and including more recently published and some unpublished cluster surveys, the Dias catalog includes a total of 1,629 objects as of its 2007 version and an extensive bibliography. The catalog merges available data to present a single table containing summary

cluster information including fundamental cluster positional and kinematic information, reddening, distance, and age. It is important to note that the catalog contains any clusters that have been identified as candidates in surveys, and exercises no selection on data quality, nor does it explore or correct for any systematic effects between the wide variety of studies it draws from. While the compilation of data into a single table is a valuable resource, the extreme heterogeneous nature of the data in the Dias catalog must be taken into consideration in the interpretation of any conclusions drawn about the overall properties of the cluster population.

Feeding the cluster catalogs are an increasing number of surveys aimed at uncovering more clusters in obscured areas of the galaxy or in areas that have not yet been systematically explored, such as the inner galaxy, the galactic plane, and regions of star formation where embedded clusters might be found. The release of the 2 Micron All Sky Survey (2MASS, available at www.ipac.caltech.edu/2mass/releases/allsky), in particular, spurred work on identifying cluster candidates that would have been hidden from previous optical surveys. While there have been many efforts in this area, two groups have been particularly active.

Bica, Dutra, Bonatto, and colleagues have published an extensive series of papers in the search for new infrared star clusters and stellar groups (e.g., Bica et al. 2003). These studies have focused on particular galactic regions where cluster surveys are known to be incomplete, where there are known optical and radio nebulae, and where 2MASS was likely to open new windows. These studies, along with detailed follow-up work, have shown that many initially identified cluster candidates were simply blended images.

Froebrich et al. (2007, FSR) carried out a systematic, automated search supplemented by a visual selection for infrared star clusters within 20° of the galactic plane using 2MASS, identifying 1,788 cluster candidates. Of these some 40% were previously known open and globular clusters, and for the remainder of new candidates, they estimated a contamination rate of 50%. This is an important caution for the many studies that are identifying clusters by their local stellar density enhancement. As the authors note, the high star density near the galactic center hampers detection, and variable star density in general introduces significant selection effects that complicate the interpretation of number statistics. Based on fitting models to the cluster density profiles, they distinguish statistically between open and globular cluster candidates and conclude that the vast majority of objects identified are expected to be open clusters.

The thorough FSR survey has spurred a number of follow-up studies (e.g., Bonatto and Bica 2008; Froebrich et al. 2008) to investigate the properties of these candidate clusters, particularly with interest in finding new globular clusters, or massive young and intermediate age clusters. These and other authors note that it is essential to have a robust means for correcting for the strong field star contamination in these cluster fields, and to utilize a variety of techniques for investigating the reality of the cluster beyond the traditional color-magnitude diagram, such as radial density profiles and mass functions. Froebrich et al. (2008) summarize the results of these follow-up studies to date, finding that of the 74 clusters investigated in more detail, approximately half of them had parameters that could be determined. Of these eight were young open clusters with ages less than 100 Myr, and half of them had ages greater than 1 Gyr. They conclude that the FSR catalog contains a large fraction of open clusters, both inside and outside the solar circle. While studies of these cluster candidates are challenging, they clearly offer a means to increase the known cluster sample in important ways and provide insight into processes of star and cluster formation and galactic structure.

This is not the place to list every survey or automated search of 2MASS or the digitized sky surveys that has been undertaken; the literature saw a plethora of them beginning in 2002. While there may be some treasures hidden in these surveys, it is important to recognize

that distinguishing between true gravitationally bound, physical associations from chance aggregates on the sky requires in-depth follow-up for individual cluster candidates. This is particularly a concern for the analysis of the increasing number of poorly populated, sparse clusters being uncovered in large scale, automated searches, and as the searches push to find the more interesting, distant, and older clusters in areas of the galaxy that have not been explored before. As several works by Carraro and Janes illustrate (e.g., Carraro et al. 2005), the large and increasing stellar densities in working toward the galactic center, coupled with the patchy and sometimes dense obscuration, result in the appearance of a distinct main sequence in color-magnitude diagrams of the field star population, simply due to geometrical effects. Patchiness in the obscuring material can create the appearance on the sky of groupings of more distant, bright stars that are interpreted as star clusters, which, in fact, have no physical association.

Selection and detection biases in the search algorithms, whether these are automated or traditional visual inspections, must also be carefully considered. Because open clusters are found in the galactic disk, background stellar density can be both high and highly variable, due to dust obscuration and to the overall features of galactic structure. These characteristics of changing background density introduce significant observational selection effects in cluster samples that naturally rely on distinguishing enhancements in cluster stellar density against the background. The ease of identifying a cluster will also depend sensitively on fundamental cluster parameters, such as intrinsic cluster richness, angular size or compactness, apparent brightness of its members, and the state of dynamical evolution of the cluster. For example, poorly populated or sparse clusters will be much more difficult to find, if at all, against a dense stellar background; those poor clusters that one does find can be expected to have systematically smaller sizes relative to clusters measured against more sparsely populated fields. Bonatto et al. (2006) offer a nice discussion of these effects. Similarly, geometrical effects due to obscuration in the galactic plane will lead to the preferential discovery of open clusters at higher galactic latitudes at greater distances; the clusters located within or closer to the galactic plane will be preferentially obscured, particularly at optical wavelengths. Because of the systematically varying stellar density with galactocentric radius and height out of the galactic plane, this introduces strong selection effects with galactic position that must be considered in interpreting the dependence of cluster properties with location.

2 Open Clusters as Stellar Laboratories

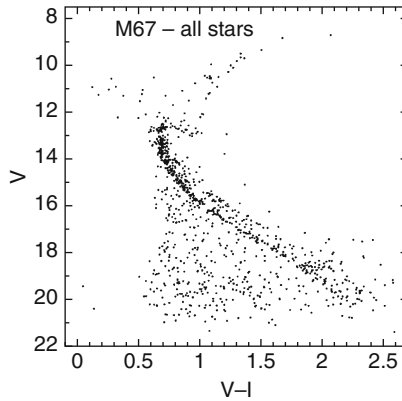
From the earliest color-magnitude diagrams of nearby open clusters, it was clear that these ensembles of from tens to hundreds of stars were valuable tools for the study of a wide range of astrophysical phenomena. The basic assumption that stars in a cluster share a common age and common chemical composition, reflecting the environment of the molecular cloud from which they formed, appears to hold from detailed observational studies. As stellar aggregates, they are valuable tools with which to study issues of stellar evolution, stellar interactions and dynamical evolution in gravitationally bound systems, star formation, and stellar nucleosynthesis.

2.1 Color-Magnitude Diagrams

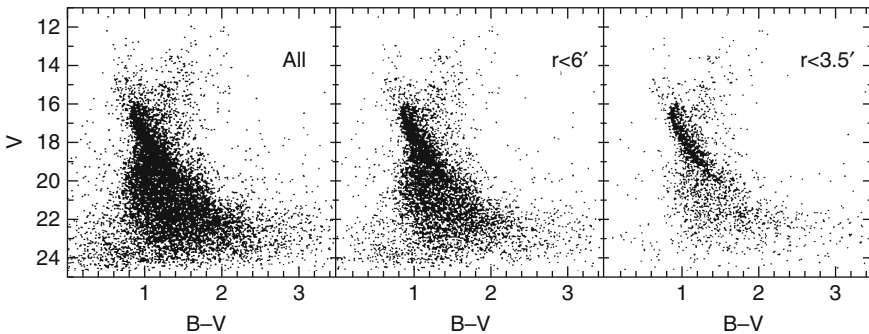
Color-magnitude diagrams of open clusters provide a fundamental tool and diagnostic, more widely used than any other for the determination of cluster properties. Color-magnitude diagrams (CMDs) reveal not only the evolutionary state of the cluster, but its stellar constituents,

such as frequency of binaries, the existence of anomalous stars, the nature of its luminosity and mass functions, and its overall properties such as reddening, distance, and metallicity. No wonder the color-magnitude diagrams are the first means with which to study cluster properties. Analysis of cluster CMDs is not without its challenges, however. [▶ Figures 7-1](#) through [▶ Fig. 7-5](#) show color-magnitude diagrams for a sample of clusters. The CMD for M67, a nearby, well-studied, 4 Gyr old cluster with solar metallicity, is shown in [▶ Fig. 7-1](#) (Montgomery et al. 1993). It traces out a clear main-sequence, well-articulated turnoff region, clearly populated subgiant and giant branch, with well-defined concentration of He-core burning “red clump” stars. Even its binary sequence stands out clearly as the stars distributed up to 0.75 magnitudes above the main sequence and in its scattering of blue stragglers brighter and bluer than the main-sequence turnoff.

Not all CMDs are so clean, however, as that for Collinder 261 show in [▶ Fig. 7-2](#) (Gozzoli et al. 1996). Its main sequence is confused with a populous field population. Only as one limits the field to the most central regions does the cluster main sequence become clearly



■ Fig. 7-1
Color-magnitude diagram for stars in the cluster M67, from Montgomery et al. (1993)



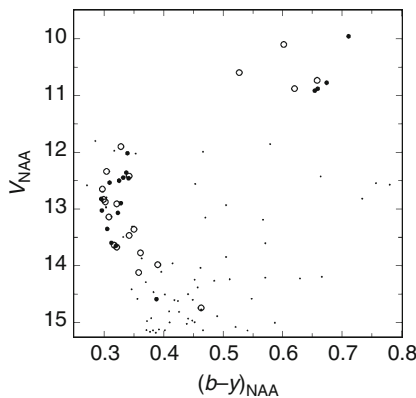
■ Fig. 7-2
Color-magnitude diagrams of Cr 261 from Gozzoli et al. (1996). Panels show the selection for the full field observed, with increasingly smaller field size centered on the cluster, to show the increasing contrast of cluster to field population

distinguished from the field, which still predominates in number. Deconvolution of the cluster from the field is extremely difficult in a case such as this, although statistical subtraction by sampling a nearby comparison field is possible. That process, too, has its limitations, as the large radii of open clusters can lead to cluster members being located at substantial distances from the cluster center, making the search for a true comparison field in strongly varying background a challenge. The solution taken by Gozzoli et al. (1996) in their analysis of Cr 261 was to add field stars to a simulation of the cluster evolutionary sequence and then compare theoretical isochrones to this synthetic CMD.

The case of NGC 3680 (Nordstrom et al. 1997, [Fig. 7-3](#)) presents the extreme of a sparse cluster in which the determination of cluster membership is critical to an understanding of its properties. Only with precise radial velocities and proper motions, which allow the determination of membership and binarity, is it possible to identify the cluster members and define the single-star evolutionary sequences of the cluster. Of the 120 stars in the field of this cluster, only 44 are cluster members, and of these, 25 are found to be binaries. Many of the stars that appear to extend the lower main sequence are not cluster members, and this careful study of membership reveals a cluster in the last stages of dissolution, its low mass stars having evaporated, leaving a severely truncated mass function.

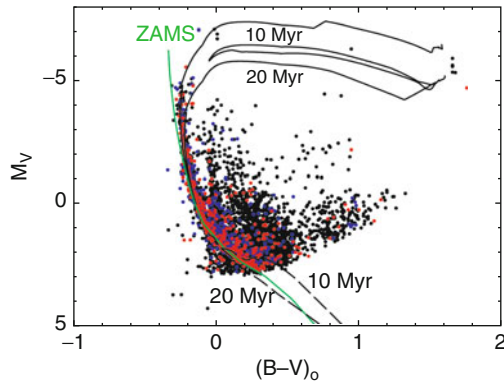
Very young clusters present their own challenges in the interpretation of color-magnitude diagrams (e.g., η and χ Persei from Slesnick et al. 2002, [Fig. 7-4](#)). In color-magnitude diagrams of young clusters, the lower main sequence appears broadened as stars approach the main sequence from the brighter luminosities. With many young clusters embedded in nebulosity, with variable reddening, and with the most massive stars already leaving the main sequence while lower mass stars are still approaching it, the interpretation of cluster membership and CMD morphology is a complex task. Multi-wavelength studies, particularly in the IR, the use of narrow band imaging, such as $H\alpha$, and spectroscopy are important tools to use in untangling membership to reveal the true cluster sequences.

Assuming that one is successful in producing a cleaned CMD for a cluster, one has a valuable diagnostic that can be fit with theoretical isochrones to derive fundamental cluster properties,



■ Fig. 7-3

Color-magnitude diagram for NGC 3680 from Nordstrom et al. (1997). *Small dots* denote nonmembers, *large dots* denote single members, and *open circles* binary members



■ Fig. 7-4

Color-magnitude diagram for stars in the field of h and χ Persei from Slesnick et al. (2002). Also shown are zero-age main sequence and post-main sequence isochrones of 10 and 20 Myr as *solid lines* and corresponding 10 and 20 Myr pre-main-sequence isochrones as *dashed lines*. *Blue and red dots* represent stars within $7'$ of the centers of h and χ Per, respectively, while *black dots* represent stars in a full $1^\circ \times 1^\circ$ field

such as reddening, distance, metallicity, and age. For many clusters, none of these parameters are independently constrained, so the traditional method of fitting theoretical sequences to the observed CMD becomes an exercise in judgment, trading goodness of fit among a matrix of models parameters, evaluated subjectively. Parameters derived from these fits are also coupled, so that one eventually obtains a self-consistent, but not unique, fit for the combined parameters of cluster age, reddening, distance, and metallicity. There have been some attempts to remove or to quantify the subjective aspect of the analysis of cluster CMDs (e.g., von Hippel et al. 2006), although the traditional approach of fitting isochrones by eye is still the predominant method.

In addition, the choice of theoretical models and their variety of assumptions about input parameters such as opacities, treatment of convection, and most importantly, their transformations from the theoretical to observational plane, result in a range of deduced cluster properties even based on the same set of observational data. The extent of these effects is demonstrated nicely by Grocholski and Sarajedini (2003) who have compared observations of a sample of clusters from the WIYN Open Cluster Survey (WOCS) to a variety of commonly used theoretical isochrones from the literature. They conclude that none of the theoretical models reproduce the observational data in a consistent way over the entire cluster main sequences, and that significant differences in isochrone shape and zero-point exist. Clearly, the deduced cluster parameters will depend on the adopted models.

Uncertainties in the range of parameters resulting from isochrone fits can be appreciable. It is not unusual to see differences of 10–20% in ages and distances, and even larger uncertainties in reddening resulting from independent analyses of a given cluster. As an example, the age of the oldest of the open clusters, Berkeley 17, has ranged from 12 Gyr (Phelps 1997) to 8.5–9.0 Gyr (Bragaglia et al. 2006). Discrepancies between studies on the order of 2–3 Gyr for the older open clusters are not uncommon depending on which models are used, how CMD morphologies are interpreted, and the emphasis placed on various aspects of the fit to evolutionary sequences.

2.2 Structural Properties and Dynamical Evolution

2.2.1 Structural Properties and Masses

Open clusters present a wide array of appearance, from sparse irregular distributions of a few dozens of stars, to rich, populous spherical concentrations. This variety of objects presents challenges in determining their structural properties observationally. Nevertheless, King (1962, 1966) and others (e.g., Mathieu 1984) have shown that modestly populated open clusters have surface brightness profiles expected for isothermal spheres modified by tidal forces. Model fits typically yield core radii of 1–2 pc, and tidal radii of 10–25 pc, although the sparse nature of open clusters, combined with their superposition on often crowded stellar fields and contamination by field stars, makes the determination of tidal radii challenging. Often the tidal radii quoted are computed from the limiting radius in the galactic tidal field assuming circular motion based on the cluster's mass. The resulting concentration ratios, defined as $\log(r_t/r_c)$ are $\lesssim 1$, by contrast to the globular clusters, which typically have concentration ratios of ~ 1 –2, an order of magnitude or more higher. There are exceptions, though, with some of the sparse globular clusters, such as AM4 or Pal 4 with concentration parameters in the range of open clusters.

The velocity distributions in open clusters are typically only a few km s^{-1} , consistent with those predicted from dynamical fits to the spatial distributions. For example, Mathieu finds a velocity dispersion of $1.2 \pm 0.35 \text{ km s}^{-1}$ for M11 and $0.25 \pm 0.18 \text{ km s}^{-1}$ for M67 (Mathieu 1985). Because of these low velocity dispersions, it is difficult to detect long period binaries, which can distort the velocity distributions by populating the high velocity tails. But once this effect is corrected for, open clusters offer an opportunity to study directly the velocity distributions as a function of mass over a wide spectrum. A recent extensive radial velocity study by Geller et al. (2008) for NGC 188 derives a global velocity dispersion of $0.64 \pm 0.04 \text{ km s}^{-1}$, which they judge may be inflated by up to 0.23 km s^{-1} due to unresolved binaries. When corrected for unresolved binaries, the radial velocity dispersion has a nearly isothermal radial distribution.

Direct determination of masses of open clusters by star counts and observed luminosity functions is challenging for both observational and physical reasons. The typical apparent cluster diameter of about 5 pc refers to a visual impression of the cluster size and, although not strictly defined as a structural property, is most closely related to a half-mass diameter. Because of the low concentration of open clusters, many studies have not sampled the full extent of the cluster. Dynamical models also predict that as clusters evolve, increasing numbers of stars will be found around the much larger tidal radii, well outside the apparent cluster diameters. The numbers of stars often attributed to clusters by simple star counts in the cluster field can then be quite misleading and result in a serious underestimate of the total cluster mass. In addition, open clusters are seen to have substantial numbers of binaries, most often detected as distinct sequences in color-magnitude diagrams, but also identified in long-term radial velocity surveys. Binary fractions may reach 50%, and mass estimates based on star counts and observed luminosity functions must be corrected for this large and often uncertain factor.

Observed luminosity functions for clusters often cover a limited range in mass, and so require a sometimes significant extrapolation of the mass function to lower masses, where much of the mass potentially resides. This effect is counterbalanced by the fact that older clusters have undergone significant dynamical evolution and preferential loss of low mass stars. The mass segregation expected, and seen, in open clusters, also complicates the determination of cluster mass, since most cluster studies concentrate on the central regions, where mass segregation has compounded the depletion of low mass stars.

Determinations of the masses of open clusters attempt to correct for many of these effects. Mathieu (1984), for example, carried out a thorough study of the structural parameters and dynamics of the 200 Myr old cluster M11. With a core radius of 0.72 pc and tidal radius of 15 pc, M11 shows significant mass segregation. After correcting for binaries, scaling counts from the observed region to the entire cluster by using dynamical models, and considering corrections for mass segregation and cluster mass outside the tidal radius, he finds a total cluster mass of $5,200 M_{\odot}$. However, this determination reaches only to masses of $0.7 M_{\odot}$, and is surely a lower limit; were the cluster mass function to follow that of the field, perhaps as much as 40% of cluster mass may be in stars with masses from 0.1 to $0.7 M_{\odot}$.

Similarly large present-day masses have also been determined for the 8–9 Gyr old cluster NGC 6791, for which Kaluzny and Udalski (1992) derive a conservative lower limit to the total cluster mass of $4,070 M_{\odot}$ from the observed stars with $V < 21$ (masses of greater than approximately $0.6 M_{\odot}$), without correction for binarity or incompleteness in the data. Geller et al. (2008) determined the virial mass for the 7 Gyr old NGC 188 to be $2,300 \pm 460 M_{\odot}$. A recent very thorough study of the young double cluster h and χ Persei by Currie et al. (2010), identifies as many as 20,000 members in the region, with a total mass of at least $20,000 M_{\odot}$. This direct determination reaches down to mid-M dwarfs, includes members in the low density halo region of the clusters, and carefully considers membership to ensure a complete census of the cluster stellar population.

Although open clusters are commonly thought of as groupings of a few dozen to a few hundred stars, it is clear that their total masses cover a wide range in the mass spectrum. There exist quite massive open clusters in the galaxy today, with present-day masses over $10^4 M_{\odot}$, and initial masses much larger.

2.2.2 Cluster Dynamical Evolution

Open clusters are strongly affected both by internal dynamical evolution and by encounters with external forces, such as molecular clouds and the galactic tidal field. These effects lead to significant modification of the cluster internal structure and mass distributions, and control the cluster longevity.

Given typical open cluster sizes (radii of $\sim 1\text{--}2$ pc) and velocity dispersions ($\sim 1 \text{ km s}^{-1}$), crossing times for open clusters are only a few Myr. Stellar encounters through these crossings modify the velocities, causing equipartition of energy between stars of different masses. The timescale for this process, or the relaxation timescale, is dependent on the number of cluster stars, as $0.1N \cdot t(\text{cross})/\ln(N)$. For a typical open cluster, with several hundred to a thousand members, the relaxation timescale is on the order of a few tens of Myr, many times the crossing time. And with typical ages of a few hundred million years, energy equipartition and mass segregation can be expected in mature, bound open clusters. The lower mass stars will have migrated to the outer regions of the cluster, while the more massive stars will have collected in the cluster centers.

This mass segregation is seen in almost all open clusters of sufficient age. Mathieu's (1984) study of M11, a 200 Myr old cluster, shows clear evidence of mass segregation; luminosity functions in the inner regions of the cluster are flatter than those in the outer regions, indicating a deficit of low mass stars. The effects of mass segregation are often most clearly seen when comparing the relative distributions of red giants and main-sequence stars, with the more massive red giants appearing more centrally concentrated than the single main sequence stars. More recent studies, able to reach to lower stellar and even substellar masses, show mass

segregation in clusters such as the Pleiades at 120 Myr, (Moraux et al. 2004), Praesepe at ~ 600 Myr (Kraus and Hillenbrand 2007), and the Hyades at ~ 700 Myr (Bouvier et al. 2008), with significant depletion of low mass stars.

The movement of low mass stars to the outer regions of the clusters and the low velocity dispersions of clusters result in the gradual evaporation of low mass stars from the cluster over time. These low mass stars collect in an outer halo beyond the tidal radius of the cluster. Independently of any external influences, open clusters will dissolve into the surrounding field. The timescale for evaporation is on the order of ~ 100 times the relaxation time, setting an upper limit to the lifetime of any bound open cluster (Spitzer 1958). For the typical open cluster of several hundred stars and relaxation times of a few tens of Myr, this results in evaporation timescales on the order of 1 Gyr or larger for more populous clusters.


As Spitzer (1958) pointed out, however, there are more important forces acting on open clusters as they orbit in the galactic potential. Both tidal forces and encounters with giant molecular clouds are extremely effective in disrupting clusters. The efficiency of disruption by clouds is an order of magnitude more important than internal effects for most open clusters, and most clusters are expected to be disrupted on timescales of a few times 10^8 years. The internal mass segregation that brings low mass stars to the outer regions of the cluster aids in the eventual disruption of the cluster, as these stars near the tidal radius are more vulnerable to tidal forces and more likely to gain enough energy to escape the cluster altogether. These internal and external disruptive factors are so effective that one does not expect to see clusters with ages greater than about 1 Gyr.

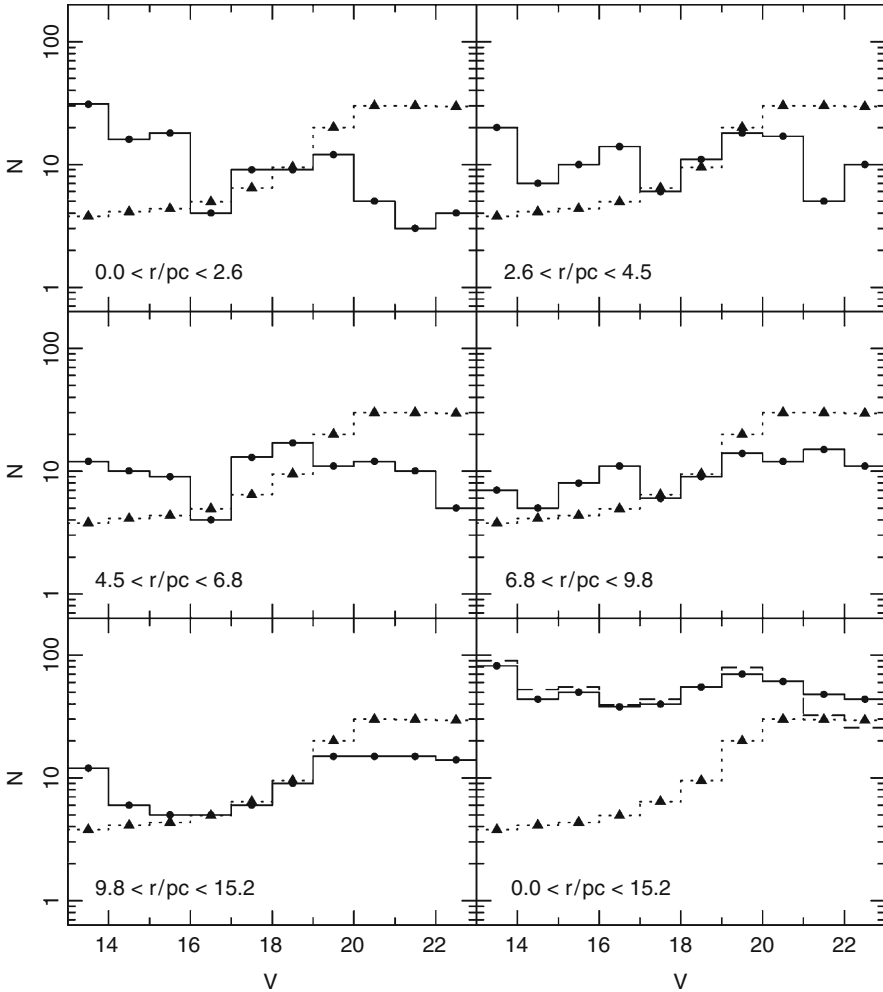
A number of N-body simulations have demonstrated, in much more detail, the mechanisms and effectiveness of both internal and externally forced cluster dynamical evolution. The relatively small number of stars in open clusters, relative to globular clusters, make them ideal subjects for direct N-body techniques and work over the years has incorporated increasingly realistic physical effects. The first extensive study of open cluster dynamics using N-body calculations was Terlevich's (1987) analysis of an $N = 1,000$ body simulation that included the effects of mass loss from stellar evolution, binary encounters, tidal perturbation from the smooth galactic field and shocks from encounters with interstellar clouds of different masses, densities, and spatial distributions. Her models show clearly the expected mass segregation, with stars at the tidal radius taking a long time to leave the cluster, where a corona of 50–80 low mass stars form between 1 and 2 tidal radii. The formation and evolution of binaries in the cluster core can have an important effect on the overall evolution of the cluster, more so for the less populous clusters. She finds that mechanisms for the disruption of the cluster depend on the cluster mean radius, and that clusters with radii of 2–3 pc have the largest lifetimes. Encounters with standard interstellar clouds are not effective at disrupting clusters, once other factors such as mass loss, binary interactions, and the smooth tidal field are taken into account. Only encounters with massive giant molecular clouds are capable of disruption of typical open clusters.

In a series of N-body studies that explored successively the impact of different initial mass functions (IMFs), mass loss through stellar evolution, and primordial binary populations, de la Fuente Marcos (1997) found that the interplay of these effects is critical. The initial mass function (IMF) plays an important role in the evolution of binaries, which in turn can strongly influence the cluster evolution. In poorly populated clusters of several hundred members, for example, massive stars in binaries can control the cluster evolution, and in small clusters can accelerate cluster disruption. For models with small N and numbers of binaries, the effects of stellar evolution are dominant in the dynamical evolution of the cluster.

In increasingly more realistic simulations, with larger number of bodies, Portegies Zwart and colleagues studied N-body models that were fully self-consistent in the treatment of stellar evolutionary effects, binary evolution, which included both dynamical and stellar effects, and tidal effects from the general galactic field, in an effort to reproduce in detail the structural and evolutionary characteristics of Hyades-like clusters (Portegies Zwart et al. 2001, 2004). With model clusters of 2,000–3,000 stars, they found that the effects of stellar evolution have more impact in the early stages, resulting in about 20% overall mass loss to the cluster. Yet the total mass loss is larger than the sum of stellar evolution and dynamical effects by some 50%, suggesting that it is the interplay between stellar evolution and dynamics that is important, especially at later times. Their simulations show that the cluster relaxation time evolves through the cluster's lifetime, implying that present-day estimates of the relaxation time may be misleading and not reflect the dynamical age of the cluster. Nevertheless, mass segregation happens early on in their simulations, in agreement with observation. In comparison to well known clusters such as the Hyades, Pleiades, and Praesepe, however, the observations show even flatter luminosity functions than the models. This could be explained if the clusters exhibited primordial mass segregation, or they had higher initial masses than the models assumed, but with shallower density profiles so that the selective evaporation of lower mass stars results in a dynamically older appearance.

Lamers and Gieles (2006) and Gieles et al. (2006) carried out N-body simulations of open clusters including encounters with molecular clouds and shocks from spiral arm passages. They found that the influence of giant molecular clouds is ten times more effective than spiral arm shocking, and accounts for as much effect as all other factors combined. The disruptive timescale for a $10^4 M_{\odot}$ cluster in the solar neighborhood is 2 Gyr. A cluster with $M \sim 10^3$ to $10^4 M_{\odot}$ is destroyed by just a few encounters with a giant molecular cloud, so individual clusters will have strongly varying lifetimes depending on their particular history of encounters.

The most complete and realistic analysis of the dynamical evolution of a single cluster has been carried out by Hurley et al. (2005) in their modeling of M67. Working from the present-day properties of M67, they use N-body simulations to model the evolution of this 4 Gyr cluster, deducing its original structural parameters and stellar populations. They find that the best fit to the cluster's current total visible mass of $1,400 M_{\odot}$ and half-mass radius of 4.3 pc is found with a cluster initial mass of $19,000 M_{\odot}$, composed of 12,000 single stars and 12,000 binaries. After its 4 Gyr of evolution, the total mass has reduced to only $2,000 M_{\odot}$ as a result of mass loss and evaporation of stars. M67 is dynamically relaxed, having passed through 13 half-mass relaxation times, and is old enough that all information about the initial mass function is lost. Structural parameters change dramatically as the cluster evolves. For example, the cluster initial core density was $150 \text{ stars pc}^{-3}$, increasing to a maximum of $330 \text{ stars pc}^{-3}$ at 3.5 Gyr, and only 83 stars pc^{-3} at its present age of 4 Gyr. The simulated cluster luminosity function shows a clear radial dependence with the central region containing many more massive stars than low mass stars, inverting the slope from its initial shape. The slope of the luminosity function becomes flatter as one moves outward in the cluster, as lower mass stars become a more dominant component of the cluster population.  Figure 7-5, from Hurley et al. (2005), shows the simulated evolution of the luminosity function in radial bins, compared to the original function without dynamical evolution. Only in the outermost regions does the slope of the present-day cluster luminosity function agree with the initial function, but even then, the cluster is depleted in the lowest mass stars.



■ Fig. 7-5

Evolution of the luminosity function of single main sequence stars in M67 from N-body simulations of Hurley et al. (2005). Panels show five radial zones from the center (*upper left*) and the cumulative distribution (*lower right*). Dotted line (with triangles) in each panel is the luminosity function expected based on population synthesis of a 4 Gyr cluster with no dynamical evolution. There is a clear radial dependence with the central regions containing many more massive stars. Even in the outer regions, the simulated luminosity function shows a deficit of low mass stars

They also follow the changing nature of the cluster's stellar population mix with time, including the formation and evolution of binary and multiple systems. The mass fraction of white dwarfs is significantly enhanced by the dynamical evolution of the cluster. Their detailed dynamical models allow them to explain the mechanisms of mass transfer or mergers by which stars or binary systems populate unusual regions in the cluster color-magnitude diagram, in very good agreement with the observations.

2.3 Cluster Mass Functions

Both observations and modeling show that open clusters experience significant modification of their mass functions through their lifetimes. As stellar laboratories, young open clusters provide among the best tools with which to study the initial mass function. Particularly compared to studies of field stars, clusters present the advantage of a coeval population at a common distance with common chemical composition, so provide an instantaneous sampling of the IMF at different locations and times in galactic history. Determining the mass function in clusters has its own challenges, however, as one must deal with and correct for incompleteness in samples particularly at the faint, low masses, contamination from cluster nonmembers, and the dynamical evolution that modifies, on a relatively rapid timescale, the mass distribution of the cluster.

As Lada and Lada (2003) point out, the use of young embedded clusters for IMF determinations alleviates many of these issues. They show, for the most well studied of all embedded clusters, the Trapezium in Orion, a present-day mass function that can be traced from $\sim 10 M_{\odot}$ OB stars to $0.1 M_{\odot}$ brown dwarfs. This mass function features a sharp power law rise from $\sim 10 M_{\odot}$ to $0.6 M_{\odot}$ with a slope of -1.2 (defined on a logarithmic mass scale where the slope = $\partial \log \xi / \partial \log m$), a slope very similar to the value of -1.35 originally derived for field stars by Salpeter (1955). At lower masses it flattens, with a slow rise to a peak at $\sim 0.1 M_{\odot}$, followed by a steep decline into the substellar, brown dwarf range. The broad peak of the IMF, which extends roughly from 0.6 to $0.1 M_{\odot}$ demonstrates that there is a characteristic mass produced by the star formation process in Orion. The IMF for the Trapezium agrees with that from field stars (Kroupa 2002), suggesting that the IMF and star formation process that produces it is very robust for stellar mass objects. The observed variation in luminosity functions in other clusters can be explained by luminosity evolution in pre-main-sequence stars of clusters of different age, but reflect a similar underlying mass function.

In exploring the question of the universality of the mass function in older clusters, understanding and compensating for the effects of dynamical evolution become important. Recent work that has probed the lowest masses in open clusters in the age range of ~ 100 – 700 Myr has provided evidence for remarkably uniform mass functions, across a range of cluster properties and environments. In a series of studies by Moraux and colleagues, comparisons of the Pleiades, Blanco 1, and the Hyades have shown mass functions in the range of 0.03 – $3 M_{\odot}$ that are fit by a common log-normal distribution (Moraux et al. 2007). When consideration is taken of the expected dynamical evolution and preferential loss of low mass stars, they conclude that the present-day mass functions of these clusters are consistent with a common initial mass function that is similar to that of the galactic field. The fact that the initial mass function does not seem to depend on the environment, from clusters of quite different masses, densities, and star-forming environments, places strong constraints on theories of star formation.

However, with the discovery and study of massive young clusters near the galactic center, such as the Arches cluster and NGC 3603, questions arose as to the universality of the initial mass function. Initial work revealed mass functions with slopes that were more shallow and surprisingly strong evidence for mass segregation in the cores of these massive young clusters (Kim et al. 2006; Stolte et al. 2006; Harayama et al. 2008). With ages of only 1 – 3 Myr, shorter than the cluster relaxation time, such strong mass segregation was not expected. Was the mass segregation primordial, and did it reveal something about the processes of high mass star formation in cloud cores?

The interpretation of the apparent mass segregation in these young clusters is not yet clear, although a variety of explanations can be found. As a number of authors pointed out, although

the cluster relaxation time is longer than the cluster age overall, dynamical evolution could have operated on the most massive stars on timescales on the order of the cluster age, resulting in mass segregation. Portegies Zwart et al. (2007), in a study of the Arches cluster, showed that the peculiarities in the mass function can be explained without resorting to primordial mass segregation, and that the Arches mass function is consistent with a Salpeter slope over $1\text{--}100 M_{\odot}$. The dynamical models can reproduce the observations if the cluster is midway through the process of core collapse. McMillan et al. (2007) have investigated models of star formation that can produce mass segregation in very young clusters that appear not old enough to be dynamically evolved, by assuming that stars form in small clumps that subsequently merge to form larger systems. Mass segregation in these smaller clumps, either initial or a result of dynamical evolution on very short timescales, is then preserved in the larger structures as they merge.

The observational challenges in the analysis of these data are severe, however, and others have noted that, even with these massive clusters, sample incompleteness, the difficulty of correcting for field star contamination, differential reddening, and crowding in these galactic fields, complicate the interpretation of luminosity and mass functions. Ascenso et al. (2009), for example, conclude that there is currently no robust way to differentiate between true mass segregation and observational effects.

Nevertheless, in a comparison of observational results for the Arches, R136, NGC 3603, and Orion, Stolte et al. (2006) conclude that the slopes of the present-day mass functions for these clusters from very different star-forming environments are in remarkable agreement, in accordance with a universal IMF slope. While all of the clusters show observational evidence for mass segregation in their cores, the present-day mass functions rapidly approach a normal IMF outside the core, and they suggest that a mass-segregated core with an extended stellar halo may be a common cluster structure in a variety of environments.

2.4 Stellar Evolution and Star Clusters

Star clusters have long been recognized and used as the optimum test cases for the confrontation of stellar evolutionary models with observation. Open clusters, through their wide range of ages and sampling of the full range of stellar masses, serve as probes of a multitude of stellar phenomena. Here is a sample of only a few of the areas they touch.

2.4.1 Convective Overshooting

Stars with masses in the range of $\sim 1\text{--}2.2 M_{\odot}$ develop a small convective core, from which convective elements overshoot the boundary between the convective inner zone and the relatively stable outer radiative zone. Overshooting effectively increases the mass of the core, extending the stellar lifetime, and is reflected in the detailed morphology of the HR diagram, particularly in the region of the main sequence turnoff. A variety of formalisms for treating convective overshooting have been developed, leading to a variety of theoretical predictions (e.g., Bertelli et al. 1985; Maeder and Meynet 1991). Open clusters with ages from roughly 700 Myr to several Gyr offer the ideal tests to constrain the appropriateness of particular models, the extent of convective overshooting, its dependence on stellar mass and other parameters, and its impact on the determination of stellar ages. A large number of observational studies from the 1990s to the present have offered these detailed comparisons (e.g., Daniel et al. 1994; Andersen et al. 1990;

VandenBerg and Stetson 2004). Although after the initial flurry of studies, it became clear that the magnitude of the effect of convective overshooting was not as dramatic as first identified, it is nevertheless clearly a factor that must be taken into account both in stellar modeling and in deriving accurate cluster ages and parameters.

2.4.2 White Dwarfs and the Initial–Final Mass Function

The availability of deep, precise color-magnitude diagrams of nearby open clusters has allowed the study of stellar populations that probe the final stages of stellar evolution for intermediate and low mass stars. Deep photometry over large areal extents in open clusters now reveals the faint white dwarf sequences in a number of open clusters of a range of ages, metallicities, and structural parameters (e.g., Kalirai et al. 2003). The resulting samples provide constraints on a variety of physical phenomena, from the cooling ages of white dwarfs, to the upper mass limit for white dwarf production, the relationship between the initial and final mass of a star, and the total amount of mass loss through stellar evolution, which in turn is a critical parameter for models of galactic chemical evolution and enrichment of the interstellar medium. Recent work by Kalirai et al. (2008), for example, shows a clear correlation between initial and final stellar mass, with more massive main sequence stars producing more massive white dwarfs, and total mass loss scaling with initial mass. The most massive stars that will form white dwarfs lose about 85% of their mass, while solar mass stars will lose only ~55% of their total mass.

2.4.3 Binary Stars and Blue Stragglers

Open clusters are excellent laboratories for the study of binary systems and their manifestation through dynamical evolution. The color-magnitude diagrams of open clusters frequently show distinct binary sequences, particularly in those nearby or relatively high latitude clusters with minimal field star contamination. Equal mass binary systems will appear 0.75 magnitudes brighter than single stars on the main sequence, while unequal mass systems will distribute in brightness between the two sequences. Studies of individual clusters indicate binary fractions of 20–50% are common, suggesting that the fraction of binaries in clusters is not very different from that of the field. The long-term monitoring program of radial velocities in open clusters carried out by Mermilliod and Mayor for the study of cluster membership, binarity, and rotational velocities, shows an overall frequency of spectroscopic binaries of 30% (Mermilliod et al. 2008).

Interestingly, older open clusters often show a population of blue stragglers, stars more luminous and bluer than the main sequence turnoff. These stars are thought to derive from normal main sequence stars that have increased in mass above a single star mass typical of the turnoff through mass transfer, mergers, or collisions in binary systems. The exact mechanism for the formation of blue stragglers is still not understood, but the open cluster systems provide excellent laboratories for confronting models of their formation and dynamical evolution with the properties of observed systems (Hurley et al. 2005). A recent study of the old open cluster NGC 188 by Mathieu and Geller (2009), for example, identifies 76% of the 21 blue stragglers in the cluster to be in binary systems, and their rotational and orbital properties suggest that most, and possibly all blue stragglers derive from multiple star systems, likely following several formation scenarios simultaneously.

2.4.4 Stellar Nucleosynthesis and Evolution

It is commonly assumed that members of open clusters share the chemical composition of the gas cloud from which they formed and so reflect the environment of their birthplace. How valid is this assumption? Is there any sign of self-enrichment in the clusters? Spectroscopic studies have shown that to high precision, derived stellar abundances among cluster members are highly uniform, with internal dispersions entirely consistent with expected measurement errors. De Silva et al.'s (2006) study of the Hyades, for example, found little or no intrinsic scatter among Hyades F–K dwarfs in a study of the heavy neutron-capture elements that are not thought to be modified during normal stellar evolution. Similar uniformity of abundance has been found in many studies of samples of brighter evolved stars in clusters. Yet relatively few studies have studied both cluster dwarfs and giants simultaneously with a common data set and analysis to investigate the extent of variations in abundance that might be due to expected evolutionary effects or unexpected systematic variations.

Samples that included both giants and dwarfs have been studied in the clusters IC 4651 (Pasquini et al. 2004), M67 (Randich et al. 2006), and the Hyades (Schuler et al. 2006, 2009). Taking into account the intricacies of stellar abundance analyses, these works find heavy elements (Fe, Ni, Cr) to be identical within the uncertainties for the evolved and unevolved stars, indicating, as expected, that the derived stellar abundances reflect the primordial composition.

For the light elements, there is mixed evidence for abundance variations due to evolutionary effects that bring nucleosynthetically processed material to the stellar surface. Unlike the case of the globular clusters, however, where stellar evolutionary effects on light elements such as C, N, O, Na, Mg, and Al are pronounced and strongly correlated, for the younger and more metal-rich open cluster stars, the effects are both more modest and, in some cases, ambiguous. For the CNO cycle, theoretical models for stars of the ages and compositions of open clusters predict some modification of surface C and N abundances in evolved stars, with depletion of C and enhancement of N, with O unchanged, and reduction in the $^{12}\text{C}/^{13}\text{C}$ isotopic ratio. Smiljanic et al. (2009) find the behavior of N, C, and $^{12}\text{C}/^{13}\text{C}$ in general agreement with predictions of first dredge up in a set of giants from ten open clusters. Schuler et al. (2009) find for the Hyades that the N and O abundances are in excellent agreement with prediction of the first dredge up, as is the isotopic ratio of $^{12}\text{C}/^{13}\text{C}$. However, the ^{12}C abundances are depleted in the giants much more than predicted by models. The cause of this additional depletion is unknown, although they suggest that it must lie outside of the CNO bi-cycle.


The situation with regard to the other light elements Na and Al is unclear. Sodium is commonly found to be enhanced in red giants in open clusters by values of +0.2 to 0.3 dex, compared both to field dwarf stars (Friel 2006; Sestito et al. 2008) and to unevolved stars in the same cluster (Pasquini et al. 2004; Schuler et al. 2009). Enhancement in Na and Al may be seen if the Ne-Na and Mg-Al nucleosynthetic cycles are active in the core regions and if the convective zone extends deep enough during the first dredge up to bring the processed material to the stellar surface. Standard models predict only slight enhancement, if any, in Na abundances, and no enhancement in Al. Schuler et al. (2009) find abundances for Na, Mg, and Al in Hyades giants to be much larger than those found in dwarfs, by amounts that exceed those predicted. Complicating the picture for Na and Al abundances are the poorly understood effects of corrections for non local thermodynamic equilibrium (non-LTE) for stars of these metallicities, masses, and evolutionary state. It is currently thought that much of this discrepancy for Na, in particular, is due to issues in the abundance analyses and the impact of unaccounted for non-LTE effects. But there is much work to be done in this area.

Open clusters have been primary tools in the effort to understand the abundances of the light element lithium, which is also produced in the Big Bang and whose presence in stellar atmospheres is a sensitive indicator of a myriad of stellar internal and evolutionary processes. The observed patterns of lithium abundance are a complex function of mass, composition, stellar rotation, and age, and continue to challenge theoretical understanding of stellar interiors and mixing processes. The subject of lithium abundances in open clusters would fill a review article itself, and the reader is referred to Pinsonneault (1997) for a recent review.

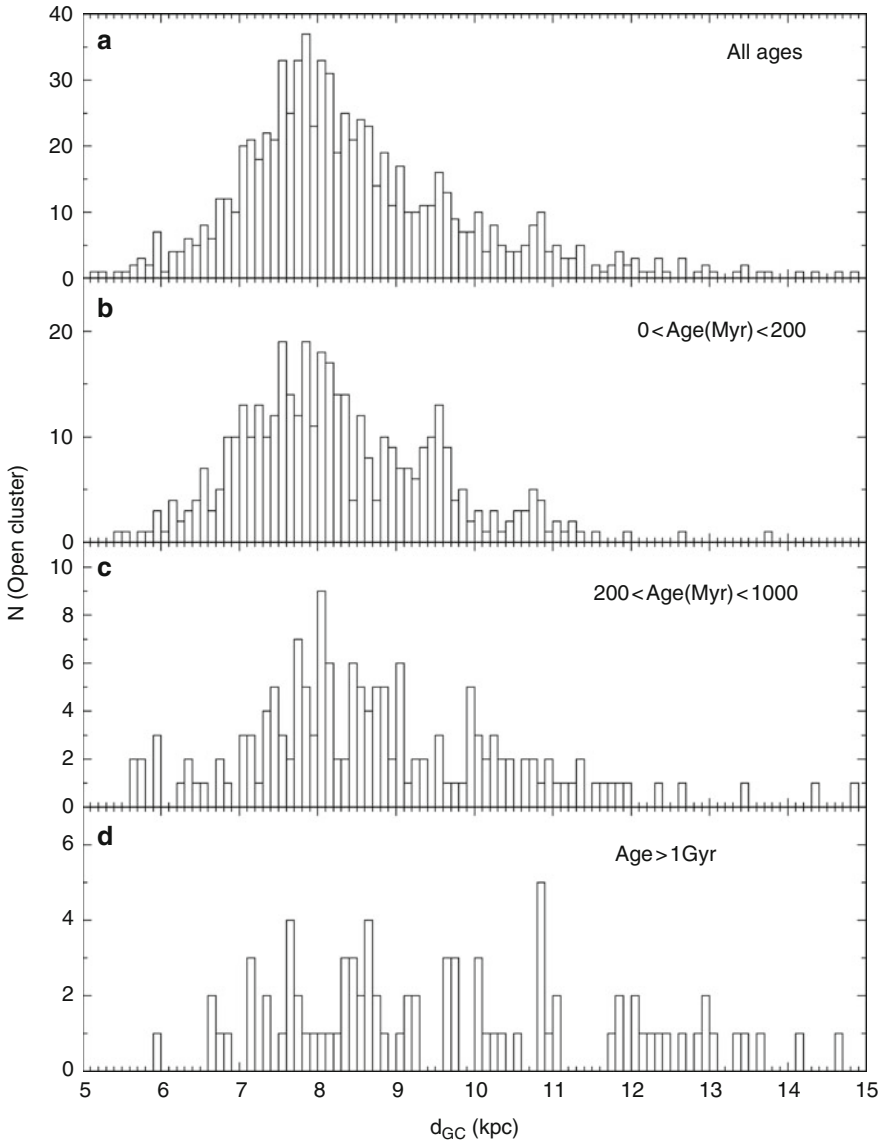
3 Open Clusters as Galactic Tracers

Open clusters can also be looked at as a system whose properties and distribution can tell us much about the processes of their formation and evolution in the context of the galaxy. Early on, even with small samples of clusters, certain characteristics were clear: The open clusters were strongly concentrated to the plane of the galaxy and many properties were correlated with cluster age and with galactic location, indicating that significant environmental processes have shaped the cluster population of today. van den Bergh (1958) noted that the four oldest open clusters were located at higher distances out of the plane than was typical for open clusters, and from a larger sample of 20 clusters, van den Bergh and McClure (1980) pointed out that clusters with ages greater than 1 Gyr were located preferentially toward the galactic anticenter compared to their younger counterparts. Oort (1958) noted that old clusters were underabundant relative to the numbers expected by extrapolating the population of young clusters and assuming a uniform rate of star formation over the lifetime of the disk. The disruptive effect of encounters with massive clouds could explain a paucity of old clusters (Spitzer 1958), but, in fact, these processes were too efficient to explain the many ancient clusters seen. The population of clusters seen today is the result of the relationships and complex interplay between processes of formation, intrinsic cluster properties, internal dynamics, and the galactic environment.

3.1 Spatial Distribution of Clusters

The first large-scale study of general cluster properties was based on the Lynga catalog (1982), as analyzed by Janes et al. (1988). A weighting scheme that considered the varying quality of catalog data, as well as a uniform and systematic approach to the determination of cluster parameters yielded a sample of 421 clusters which revealed many overall properties of the open cluster system. The first of these was the general spatial distribution of open clusters, showing clearly that its characteristics were a function of cluster age. This distinction was manifest clearly in the radial distribution, which showed that older clusters are not found in the inner part of the galaxy. From about the solar circle outward, although clusters of all ages could be found, the average cluster age increased, and older clusters predominated in the outer disk, even after considering the incompleteness effects expected at larger distances (see  Fig. 7-6).


The vertical distribution of open clusters also showed a clear dependence on age. For clusters with ages less than about 300 Myr, the exponential scaleheight from the plane was constant, at 55 pc, but increased gradually for older clusters. The distribution of these young clusters also showed a clear displacement of the sun of 21 pc with respect to the galactic mid-plane, similar to what is seen relative to field stars. However, for clusters of all ages, the disk becomes



■ Fig. 7-6

Distribution of open clusters with distance from the galactic center from Bonatto et al. (2006). Cluster distributions are grouped by age, as shown (Cluster parameters are adopted from the WEBDA database)

thicker with increasing galactocentric distance, a trend which is much more pronounced for the older clusters. Later studies quantified this dependence, when larger samples of clusters were available, particularly at the larger ages. Janes and Phelps (1994) found that the vertical distribution of clusters with ages greater than about the Hyades was well fit by an exponential with a scaleheight of 375 pc.

Twenty years later, with a cluster sample increased by 50%, and improved parameters for many clusters, Bonatto et al. (2006), carried out a similar analysis. Using a sample of 654 clusters with known distance, drawn almost exclusively from the WEBDA, they looked in detail at the dependences of scaleheight on age. Their results reinforce the general trends seen in the earlier papers. The entire cluster population shows an exponential scaleheight of 57 pc, but this value increases with increasing cluster age: Clusters with ages younger than 200 Myr show an exponential scaleheight of 48 pc, for those with ages from 200 Myr to 1 Gyr it increases to 150 pc, while clusters with ages greater than 1 Gyr are nearly uniformly distributed within 400 pc of the plane (their  Fig. 7-10). For the younger clusters their determinations agree well with other work, but for the oldest clusters, the lack of exponential falloff with height contrasts with that found by Janes and Phelps (1994). This difference may be due to the somewhat different age range considered, as the Janes and Phelps sample included clusters as young as 700 Myr, and the additional old clusters in the later work. Nevertheless, the oldest clusters are clearly found in a more spatially extended vertical distribution than the younger clusters.

The more recent, larger cluster sample used by Bonatto et al. (2006) also illustrates the thickening of the disk with increased galactocentric distance. The galactic disk as traced by the clusters with ages less than 1 Gyr shows a thickening by a factor of 2 in moving from inside to outside the solar circle. For $R_{gc} < 8$ kpc, the exponential scaleheight is $z(h) = 39.3 \pm 3.3$ pc, while outside this distance, $z(h) = 78.1 \pm 5.9$ pc. Both the values of the scaleheight and the thickening with distance are consistent with the disk as defined by H I and reflect the association of star formation with the parent gas (J. M. Dickey, this volume).

In all of these studies which rely on samples of clusters ranging over the full extent of the galactic disk, one must be aware of observational selection effects and incompleteness which can severely affect the conclusions based on observed distributions. Clearly incompleteness is a function of cluster distance. But the varying stellar density in different galactic directions will also affect the ability to distinguish clusters, with greatest impact for the discovery of sparse clusters against the dense stellar fields toward the galactic center. Bonatto et al. (2006) attempt to quantify and correct for incompleteness in their sample by using the variation in the background density from 2MASS to define the detectability of clusters projected toward different regions of the sky. Based on this model, they correct the observed radial and vertical distributions of clusters. When corrected for incompleteness, the observed cluster scaleheights increased by factors of roughly 50%.

More important is the effect the corrections have on the radial distribution of clusters and the interpretation of its general characteristics. As mentioned above, the general observation that old clusters are not found much inside the solar circle has been known for some time. This observation is clearly revealing cluster and galactic dynamics at work, but how much is observational selection? Janes and Phelps (1994), with the first substantial sample of old clusters, noted that their galactic distribution is highly asymmetric, with none found inside a galactocentric distance of 7.5 kpc (assuming a solar galactocentric distance of 8.5 kpc). Considering that the younger clusters show a distribution centered on the sun, with appreciable numbers in the inner galaxy, they argue that the distribution for old clusters cannot be entirely observational selection effects, although those must certainly come into play to some degree.

Bonatto et al. (2006) look at the radial profile of the full cluster population, which, as expected due to observational selection effects, shows a maximum at the solar position, and then decreases in number both interior and exterior to the solar circle. How much of this fall off is due to incompleteness in the sample and how much reflects the underlying number distribution of open clusters in the galactic disk? Limiting to the region within 1.3 kpc of the sun,

to minimize the effect of incompleteness, but still correcting for it, they find a disk scale length of 1.4 ± 0.2 kpc. Considering the full extent of the cluster sample, again corrected for incompleteness, the distribution is fit with a scale length of from 1.5 to 1.9 kpc. Both of these determinations reveal a disk scale length that is shorter by a factor of 2 than that derived from stellar populations (see Churchwell and Benjamin, this volume). It is not clear whether this reflects an intrinsic difference with respect to the field star population, or uncertainties in determination.

The observed lack of older clusters in the inner parts of the galaxy remains in the Bonatto et al. (2006) analysis, even after the correction for incompleteness and selection effects.

3.2 Cluster Physical Parameters

Cluster structural parameters and size, and any correlations of structural parameters with location in the Galaxy, provide important constraints to our understanding of both the processes of cluster formation and their dynamical evolution. Cluster sizes may reflect a primordial dependence on gas densities or thresholds for star formation, which vary with location in the galaxy. Alternatively, at the most basic level, the dynamical evolution of an open cluster of a given mass is determined by its linear size. A massive but small cluster will dissolve due to internal interactions, while a large cluster of the same mass will be more affected by external tidal interactions with the galactic field or molecular clouds. The distribution of cluster diameters thus sheds light on these processes.

Janes et al. (1988), and more recently van den Bergh (2006) have looked into the distribution of cluster diameters based on the catalogs of Lynga (1982) and Dias et al. (2002), respectively. From samples on the order of 400 and 600 clusters, respectively, there are clear correlations of cluster size with both location in the galaxy and with age. The largest clusters are found at the youngest ages. The majority of these extremely large clusters are unbound systems, and radial velocity studies have indicated that these associations will dissolve in a few million years. Van den Bergh estimates that approximately 20% of the “clusters” in these catalogs with ages less than 15 Myr are expanding stellar associations, rather than bound, stable clusters. In their review on embedded young clusters, Lada and Lada (2003) note that when emerging from their natal clouds young embedded clusters expand significantly and for long periods before they reach a final equilibrium. For clusters with ages less than 10 Myr, bound and unbound emerging clusters are indistinguishable.

Janes et al. (1988) found that for clusters between ~ 50 Myr and 1 Gyr, there is no correlation of cluster size with age, and conclude that there is no preferred mass scale for survivability. Looking at a somewhat finer resolution in age, and with a larger sample, van den Bergh observes that there is evidence for a slight increase in cluster size with age. Clusters with ages between 150 Myr and 1.5 Gyr are systematically larger than clusters with ages from 15 to 150 Myr. The typical cluster diameter increases from 2 pc to 3 pc in this age range. He suggests that this effect may be due to the loss of gas by the evolving stars in the cluster.

In all studies, the oldest clusters, those with ages greater than 1–1.5 Gyr, are systematically larger than the younger bound clusters. This fact may initially seem surprising; one might expect the oldest clusters to be more tightly bound to have survived to such great ages. However, these clusters have other properties that distinguish them from the majority of the open cluster population. They are also located preferentially at larger distances from the galactic plane and at large galactocentric distances in the outer disk, clearly a feature that allows them to survive.

There is also some indication that cluster diameters increase with increasing distance from the galactic center. These large samples show evidence that the proportion of clusters with small diameters is smaller at larger distances, or that larger clusters predominate at larger galactocentric distances. One might be tempted to interpret this correlation as indicating that large clusters are preferentially destroyed at small galactocentric radii or that they can survive at larger galactocentric radii because of the less frequent interactions with molecular clouds. However, this apparent correlation is strongly influenced by a variety of observational selection effects that complicate its interpretation. Small clusters will not be discovered at the same rate as larger clusters at great distances. As discussed earlier, the changing stellar background density affects the likelihood of finding sparser clusters. Most importantly, these measures are of apparent cluster diameter, not a physical parameter such as half-light or half-mass radius, so the determinations themselves are affected by the cluster mass. More populous clusters, or those seen against a less dense stellar background, will be traced to larger radii simply because there are more stars to measure or they are easier to distinguish against the low density background.

The discovery of open clusters from IR surveys, particularly 2MASS, has both increased the number of clusters over these earlier samples and allowed us to probe regions of the galaxy that may help understand the impact of dynamical effects. In a series of papers following up on the survey by Froebrich et al. (2007), Bonatto and Bica have shown that these newly discovered clusters have radii that are systematically smaller than previously known open clusters of similar age (e.g., Bonatto and Bica 2007, 2008). For clusters located inside the solar circle, the systematically smaller core and limiting radii relative to clusters outside the solar circle point to the influence of tidal effects that may have accelerated dynamical evolution.

They also find, as in earlier studies, that sizes of open clusters appear to increase with galactocentric distance, but that the newly discovered clusters tend to be smaller than previously known open clusters at the same galactocentric distance, particularly in the region from 8 to 10 kpc. As they note, part of this relation of increasing size with galactocentric distance may be primordial, and reflect the fact that the higher density of molecular gas in the inner galactic regions may have produced clusters with smaller initial radii.

Among the FSR clusters, they find a rough trend of increasing size with height from the galactic plane. This is not unexpected from dynamical effects, as clusters closer to the galactic plane suffer more frequent encounters with molecular clouds, and will survive only if they are more compact. Again, though, one must caution that observational selection effects will naturally lead to the measurement of larger cluster diameters against the lower stellar background density of the higher latitude fields.

These global correlations suggest that dynamical effects play an important role in influencing cluster intrinsic properties, as would be expected. Larger, sparser clusters are able to survive to greater ages in the areas of the galaxy less populated by large molecular clouds or other strong gravitational forces, that is in the outer disk and farther from the galactic plane. It is difficult to go beyond these general observations however, to investigate the details of the impact that overall galactic properties have on cluster evolution and longevity, because of the importance of basic observational selection effects on both the detectability of the clusters and their measured properties. Obtaining measurements of physically based structural parameters, such as the half-light or half-mass radius, rather than an apparent linear diameter, derived from a uniform data set and for a large number of clusters would undoubtedly allow progress to be made in this area. In the meantime, as discussed earlier, a great deal of insight into the overall effects of dynamical evolution on cluster structure and longevity can be obtained from the detailed modeling of n -body simulations in a realistic galactic potential.

3.3 Spiral Arms

The positions of galactic clusters and stellar associations have been used to attempt to trace out regions of star formation and spiral arms in the galactic disk. The first large samples of clusters, however, showed no clear association with the larger structure of spiral arms. Instead, young clusters (less than 20 Myr) clumped in several large complexes, not aligned with spiral structure, while older clusters showed no distinguishing distribution (Lynga 1982).

The larger optical samples recently analyzed by Bonatto and Bica (2006) and van den Bergh (2006), show stronger evidence of the longitudinal distribution of young clusters reflecting areas of active star formation and the presence of spiral structure. Enhancements in the number of clusters are seen at $l \sim 285^\circ$, coinciding with the Carina arm, particularly in the youngest clusters with ages less than ~ 10 Myr, and at $l \sim 125^\circ$ coinciding with Cassiopeia. When projected onto the galactic plane, clusters with ages less than 60 Myr show a distribution coincident with the Orion spiral arm, and a lack of clusters in the interarm region between the Orion and Sagittarius arms, just inside the solar position. Beyond this immediate solar region, the detailed distribution of young clusters shows no particular pattern suggestive of association with spiral structure.

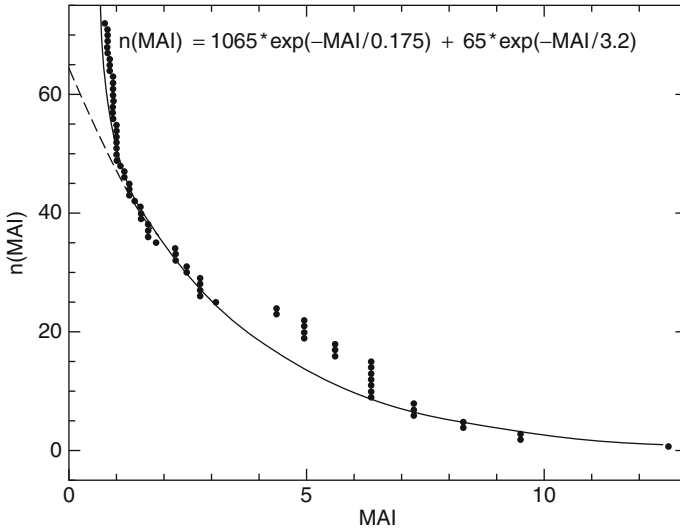
New infrared surveys, however, with their ability to probe areas of high extinction and active star formation provide an opportunity to uncover clusters either in the process of formation or emerging from their natal clouds, whose galactic distribution may indicate spiral structures.

The Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE) project has mapped out the inner galaxy at mid-IR wavelengths (3.6, 4.5, 5.8, and 8.0 μm) and produced a database from which candidate star clusters have been detected by Mercer et al. (2005). GLIMPSE data are very effective at discovering heavily embedded young clusters and their galactic distribution shows interesting asymmetries. Within their survey area they detect more than twice as many new clusters in the southern half of the Galaxy as the northern half, and more below the galactic mid-plane than above. Both of these asymmetries are seen in the optical and 2MASS cluster samples, although not as pronounced. The longitudinal distribution of GLIMPSE clusters shows distinct peaks at $l = 330^\circ$ and 313° , which may trace the location of spiral arm tangencies; the enhancement at $l = 330^\circ$ is also seen in 2MASS. The use of these surveys to explore the galactic distribution of newly discovered clusters is just beginning, as the detailed properties of these cluster candidates are still to be explored. They promise, however, to yield particular insight into questions of star and cluster formation in the inner galactic disk.

3.4 Longevity of Open Clusters

The dependence of spatial distribution on age that is apparent in the open cluster system points to a complex interplay between cluster formation and survivability. The distribution of cluster ages sheds further light on this balance. If the cluster population seen today were a result of a uniform rate of cluster formation combined with an exponentially declining dissolution rate, one would expect to see a simple exponential distribution with a characteristic single lifetime. What is seen is much different. The age distribution of clusters shows three distinct populations with different timescales of longevity.

The youngest population of clusters identified is one with lifetimes of only a few tens of millions of years. This group of clusters is apparent in any large sample, and their numbers rapidly decrease with age beyond a few tens of millions of years. These are stellar associations



■ Fig. 7-7

Cumulative distribution of numbers of clusters with ages greater than the morphological age indicator, MAI, from Janes and Phelps (1994). The solid line shows the function with exponential timescales of 175 Myr and 3.2 Gyr. The dashed line shows only the second, longer decay timescale appropriate for the older clusters in the distribution

that are not gravitationally bound, and are in the process of dissolving and dispersing into the general field population.

The majority cluster population has a characteristic lifetime of a few hundred million years and presents a rather homogeneous group. These clusters dominate the galactic system of open clusters and largely define the properties of the “typical” open cluster. Janes and Phelps (1994) find an exponential decay time for these clusters of 175–230 Myr (► Fig. 7-7). Bonatto et al. (2006) derive an exponential of 123 Myr for the majority population.

As was first apparent in Janes et al. (1988) and reinforced with larger samples in Janes and Phelps (1994), the age distribution shows a long tail to larger ages, in a distribution that cannot be fit with a single exponential with a decay time of a few hundred million years (● Fig. 7-7). This group, only a few percent of the total cluster population, can be characterized with lifetimes of 3–4 Gyrs, and includes members with ages approaching the age of the galactic disk and the youngest of the globular clusters. Janes and Phelps (1994) fit this older population in the cluster age distribution with an exponential of 3–5 Gyrs. Bonatto et al. (2006) derive an exponential timescale of 2.4 ± 1 Gyr for this old population. In either case, there is a substantial population of old clusters not explained by the simple picture of uniform formation combined with dissolution.

The mix and relative numbers of these populations vary with location in the Galaxy, leading to the spatial distributions correlating with age. The outer disk clusters are systematically longer lived; the typical open cluster in the outer disk will survive about twice as long as one in the inner disk. The characteristic lifetime of open clusters of all types is a distinct function of galactocentric radius.



3.5 The Oldest Open Clusters

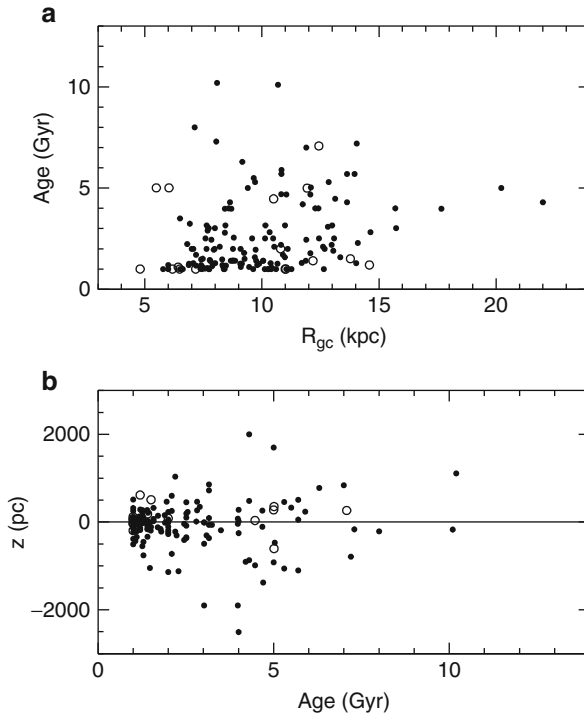
The oldest of the open clusters are clearly a special population whose detailed properties relate to a number of issues in galactic structure and formation, the dynamical evolution of clusters, and the balance between the mechanisms of cluster formation and dissolution. It is observed that clusters with ages greater than about 1 Gyr are located preferentially in the outer galaxy, almost exclusively outside about 6 kpc from the galactic center. They are also preferentially located at greater heights from the galactic plane. They are, in general, slightly larger in apparent linear diameter than their younger siblings.

Is there anything about the details of the distribution of their properties that would provide insight into what allows these clusters to survive so long?

The value of old open clusters in exploring the development and evolution of the galactic disk led, over the past several decades, to special efforts to discover old clusters through systematic searches and to determine their properties through follow-up photometry and spectroscopy. The first substantial increase in the sample of old clusters came with a survey by Phelps et al. (1994) and Janes and Phelps (1994), who identified a sample of 72 clusters with ages greater than the Hyades, or about 700 Myr. In the years following, continued effort to expand the number of old clusters known through studies of promising individual clusters, as well as the automated searches from IR surveys, have revealed increasing numbers of old cluster candidates. The follow-up studies of cluster candidates from the FSR survey, for example, have, to date, identified 16 more confirmed open clusters with ages greater than 1 Gyr (Froeblich et al. 2008).

The current Dias catalog, the WEBDA, and earlier lists of old clusters (Friel 1995) combined with follow-up of the FSR survey, reveals over 160 clusters with ages greater than 1 Gyr. This can be contrasted with the 86 clusters in the Bonatto et al. (2006) sample drawn from the WEBDA. While these numbers suggest a rapid increase in sample size, it is important to recognize that some fraction of these old ages may be incorrect estimates and that some of the “clusters” identified are simply asterisms detected in automated searches and not true bound clusters. Most importantly, the age estimates come from a variety of techniques and individual determinations, so are not on a uniform and consistent scale. Even given these caveats, with likely contamination at the level of 10–20%, the increase in numbers indicates that there are many more old clusters available from which to probe the early galactic disk.

These larger samples confirm the initial conclusions of Janes and Phelps (1994), that among these old clusters there is no correlation of age with location.  Figure 7-8 shows the distribution of cluster location with age from these heterogeneous samples, for clusters with age greater than 1 Gyr. While relative to the majority of the open cluster population with ages of several hundred million years, the old clusters are found preferentially in the outer disk and at larger heights from the galactic plane, when just the old clusters are considered, there appear to be no strong trends of age with location. As  Fig. 7-8 shows, the oldest clusters in the group, those with ages greater than ~4 Gyr, are located at all galactocentric radii. Similarly, the oldest clusters are found at all heights from the plane; they are not at preferentially larger distances than clusters with ages of only 1–2 Gyr. This observation is somewhat surprising; one might have expected the very oldest of the clusters to be the most distant or farthest from the plane in positions that have allowed them to survive the disruptive tidal forces in the plane. Yet two of the oldest clusters known, Be 17 and Cr 261, at 10 and 8 Gyr, respectively, are found only about 200 pc from the plane. Cr 261 and NGC 6791 are found just inside the solar circle, at galactocentric radii of



■ Fig. 7-8

Distribution of cluster location with age for open clusters older than 1 Gyr. Filled circles are clusters with data drawn primarily from the WEBDA, with ages from Salaris et al. (2004) when available; open circles are confirmed open clusters identified in FSR and characterized in follow-up studies (Froebrich et al. 2008). (a) Age as a function of galactocentric radius where the sun is at 8 kpc. (b) Distance from the galactic plane as a function of cluster age

7 and 8 kpc, respectively. And, the most distant open clusters known, Be 29 and Saurer 1, are only 3–5 Gyr old. All of these examples illustrate the large scatter seen among cluster properties with age.

Among these old clusters, there is a tendency for the most distant objects to be farthest from the galactic plane. This is not surprising and is expected purely from observational selection effects. Clusters at the very large distances in the outer disk are likely to be found only if they are far from the plane; those in the plane are heavily obscured.


There is much less information about the kinematics or orbital characteristics of these old clusters that might reveal more detail about their origins or reasons for longevity. However, what information there is suggests no strong correlations of kinematic behavior with age or location. Based on a sample of 35 clusters with ages greater than 1 Gyr, Scott et al. (1995) showed that the radial velocities of the old clusters are consistent with those expected from the disk rotation curve defined by the younger clusters, but with a larger dispersion of 29 km s^{-1} versus 10 km s^{-1} typical of younger clusters. Alternatively, assuming the clusters rotate about the galactic center with constant rotation velocity, the old cluster system reflected in this sample rotates with a

velocity of $211 \pm 7 \text{ km s}^{-1}$ with a line-of-sight dispersion of 28 km s^{-1} , lagging only slightly the solar rotation. These kinematics are consistent with those of the old, mixed-age, thin disk field population.

For the limited number of clusters for which one has full space motions and can compute orbits, one finds generally the same result. Orbits for five classic old clusters were modeled by Carraro and Chiosi (1994) who found eccentricities of 0.15 or less, consistent with their association with the old, thin disk population. However, the clusters are also found to be close currently to their maximum excursion from the galactic plane, where they spend most of their time away from the disk and its disruptive influences. On the other hand, there are several clusters with clearly anomalous orbits, even judged solely by their radial velocities. The cluster Be 17, the oldest open cluster known in the galaxy, which is located almost directly in the anticenter, at $l = 176^\circ$, has a radial velocity of -84 km s^{-1} (Scott et al. 1995), indicating a significantly noncircular orbit.

The old cluster NGC 6791, for which proper motions exist, has the most eccentric and unusual orbit known for an open cluster (Bedin et al. 2006). With an eccentricity of ~ 0.5 , its orbit takes it to a perigalacticon of $\sim 3 \text{ kpc}$, into a region of the galaxy where no old open clusters are currently seen, while its apogalacticon is not far beyond the solar radius. The cluster, however, has suffered strong dynamical effects due to its numerous passages through the galactic plane: In each orbital period of $\sim 130 \text{ Myr}$, it passes three times through the galactic plane, once at $R_{gc} \sim 9 \text{ kpc}$, and twice at $R_{gc} \sim 5 \text{ kpc}$, where the disk is much denser. NGC 6791 is also one of the most massive and dense open clusters, properties which are clearly responsible for its longevity.

One can ask if there is structure in the age distribution of these oldest clusters that could point to preferred periods for star formation or provide insight into the nature of the formation or disruption processes that have allowed these clusters to survive for so much longer than the typical open cluster. In looking at the subtleties of the age distribution, it is extremely important to have ages on a uniform scale, determined in similar fashion and accurate in a relative sense even if the zero-point may be uncertain. Two studies have paid particular attention to this issue.

Janes and Phelps (1994) used an age indicator based on the morphology of the cluster color-magnitude diagram, using the relative locations of the main sequence turnoff and the red giant branch and He-core burning red clump stars, to determine a “morphological age indicator” (MAI) which is well correlated with the logarithm of cluster ages. Using this measure of cluster age, they obtained a sample of 72 clusters with ages greater than the Hyades, which at the time was several times larger than had been known previously. Within this sample, they noted the possibility of an excess of clusters in the age range of 5–7 Gyr (see  Fig. 7-7), suggesting either large bursts of star formation at this period or perhaps that a larger proportion of clusters forming at that time had orbits that allowed them to survive. They argue that these old clusters cannot be just the long-lived tail of the general open cluster population because there are far too many to be explained if extrapolated from the current numbers of younger clusters. Similarly the old clusters cannot have diffused away from the galactic plane through encounters with giant molecular clouds or other massive objects, because these encounters would tend to disrupt the clusters on shorter timescales than their ages. They conclude that these old clusters must have formed from a distinct process of disturbances to the galactic disk, through infalling material or tidal interaction.

Salaris et al. (2004) further calibrated these morphological indicators in terms of absolute ages, based on a consistent set of updated stellar models, and tied to a similar set used to date a large sample of globular clusters. They also find that the distribution of these ages deviates from the simplest case of a uniform formation rate and exponentially declining dissolution

rate. They also find an excess of clusters in the 4–6 Gyr range at the 2 sigma level. This distribution is independent of galactocentric distance, which suggests that the cluster formation and destruction processes are not correlated with R_{gc} for these old clusters. The same is not the case for the z -distribution. Clusters closer to the plane follow more closely the scaling relationship corresponding to a dissolution timescale of 2.5 Gyr. The clusters more distant from the plane ($|z| > 300$ pc) show a clear excess of clusters in the range of 3–6 Gyr with a difference that is significant at more than the 3 sigma level. They suggest this points to a more homogeneous or uniform “creation–destruction process” for the clusters closer to the galactic plane. While this result reinforces the idea that the clusters survive most readily at higher distances from the galactic plane, the impact of incompleteness in the sample and selection effects at the lowest z -distances needs to be considered.

The very oldest of the open clusters can be used to set limits to the age of the galactic disk and the relative timescale of disk and halo formation. Although there are still debates on the exact ages of particular clusters, there is general agreement that Be 17, NGC 6791, and Cr 261 are the oldest open clusters currently known, with ages of 8–10 Gyrs (e.g., Phelps 1997; Bragaglia and Tosi 2006; Krusberg and Chaboyer 2006). These ages are consistent with the age of the galactic disk as probed by the white dwarfs. Salaris et al. (2004) who have determined the ages of open and globular clusters in a similar homogenous fashion determine a delay of 2.0 ± 1.5 Gyr between the start of the formation of the halo, as indicated by the oldest metal-poor globular clusters, and the thin disk formation, indicated by the age of the oldest open clusters. They also find no age difference between the thin and thick disk formation as reflected in the cluster populations. In contrast, probing the time delay between the formation of the thin and thick disks from a sample of clusters with self-consistently determined relative ages, Krusberg and Chaboyer (2006) find an age difference of 2.8 ± 0.8 Gyr between the metal-rich thick disk globular clusters and the oldest open clusters.

With many new candidate clusters with ages greater than 1 Gyr being identified in automated IR searches, it would be interesting to place them on a uniform age scale to explore further both the details of structure in the age distribution of open clusters and its relation to the globular cluster systems.

4 Galactic Chemical Evolution

Open clusters, found at all ages and throughout the galactic disk, serve as excellent tracers of the overall chemical enrichment of the disk. As they preserve the abundances of the gas from which they formed, studying the chemical profiles of clusters of different ages and locations provides a history of star formation and nucleosynthesis throughout the galaxy and across its lifetime. Recent high precision analyses indicate that stellar abundances among cluster members are highly uniform, consistent with no intrinsic scatter within the cluster (De Silva et al. 2006), so one can be assured that the abundances determined reflect the initial composition in a well-mixed gas cloud. As noted earlier, stellar evolutionary effects and mixing of processed material to the stellar surface may alter abundances of some light elements (e.g., Li, C, N) in the atmospheres of giant stars. While observations of some elements present challenges to theoretical models, they do not alter the usefulness of stellar and cluster abundances as tracers of overall galactic chemical evolutionary patterns.

Stellar abundances and overall metallicities in clusters have been derived by a wide variety of photometric and spectroscopic methods. Early work relied on photometric abundance indicators using UBV, DDO, Washington, and Stromgren photometric systems, which were able to provide overall metallicities for substantial numbers of stars in a cluster field. While photometric studies have the advantage of reaching large samples and faint, distant objects, they suffer from potential contamination by nonmembers and reddening. Significant numbers of spectroscopic studies began to appear in the 1980s, but samples were initially limited, and high-resolution spectroscopy was rare. Photometric and low-resolution spectroscopic abundances generally yield overall metallicities that are internally accurate to 0.1–0.15 dex in $[M/H]$,¹ but with the possibility of significant systematic differences between studies. Photometric and spectroscopic indices measure the blanketing in either narrow or broad bands due primarily to blends of Fe and Fe-peak elements, or common molecular species, such as CN or MgH, which are then calibrated to $[Fe/H]$ or, in the case of some photometric indices, interpreted through the analysis of color-magnitude diagrams with the help of theoretical isochrones. The Lynga, Dias, and Mermilliod catalogs of cluster parameters include metallicity determinations drawn from a wide variety of sources, and care must be taken to assess the systematic differences inherent in these collections. Systematic differences of 0.1–0.2 dex are common and can either obscure or create and certainly confuse correlations and general trends that have the potential to reveal much about galactic structure and evolution.

The availability of stellar abundances based on a uniform treatment and analysis for large samples of clusters is critical for an accurate and thorough understanding of abundance patterns both within the cluster population and on galactic scales. A plethora of differences in how measurements are made and analyzed, such as continuum tracing, equivalent widths, or spectral synthesis, or methods of analysis, such as choice of model atmospheres, adopted atomic parameters, temperature scales, or treatment of convection or assumptions of local thermodynamic equilibrium (LTE), all have a bearing on final results, and can introduce substantial systematic differences between studies. Fortunately, a growing number of studies are providing both the sample sizes, and the means to intercompare studies to understand the nature and magnitude of systematic differences that can impact the interpretation of observational results. The increasing number of high-resolution spectroscopic studies, in addition, are providing not only overall metallicity, but individual elemental profiles that shed light on detailed nucleosynthetic and star formation histories among the diverse cluster population.

4.1 Disk Abundance Gradients

The variation of abundance with position in the galactic disk provides essential constraints to models of chemical evolution, and open clusters, with their range of ages and distances are among the best indicators of trends in overall metallicity as well as abundance patterns. From the first studies it was clear that the cluster population shows a significant decrease in metallicity with increasing galactocentric distance; clusters in the outer disk are distinctly more metal-poor than those in the solar neighborhood. The exact shape of this distribution and the magnitude of the trend and whether or not it varies with age, however, are still subjects of debate.

¹Stellar abundances are defined on a logarithmic scale relative to the solar abundance, with $[X/H] = \log(X/H)_{\text{star}} - \log(X/H)_{\odot}$, where X is the number density of the element. Here M refers to an overall metallicity of heavy elements rather than an individual element.

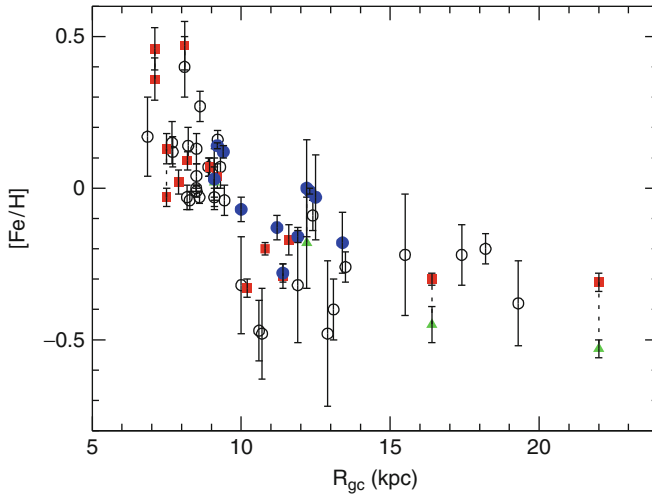
Janes (1979) offered the first quantitative description of the abundance gradient from open clusters, using a sample of 41 clusters with overall metallicities derived from a combination of DDO and UBV photometry of cluster giants. He found a gradient of $d[\text{Fe}/\text{H}]/dR = -0.05 \text{ dex kpc}^{-1}$ over a range of 8–14 kpc from the galactic center (assuming a solar distance of 10 kpc), by combining the cluster data with field red giant stars. The clusters alone indicated gradients of -0.039 ± 0.031 for clusters younger than 800 Myr, and a steeper value of -0.076 ± 0.028 for those older. Janes et al. (1988) expanded this work using the much larger sample of clusters in the Lynga catalog to derive a gradient of -0.07 for clusters younger than 200 Myr. Older clusters yielded a steeper gradient of $-0.14 \text{ dex kpc}^{-1}$, but this result was dominated by a handful of clusters in the outer part of the galaxy with metallicities $[\text{Fe}/\text{H}]$ of -0.5 to -0.6 dex. In all cases, while the overall trend of decreasing metallicity with distance was clear, the dispersion in abundance at any distance was appreciable. In neither case was there evidence for a gradient in abundance vertically from the galactic plane.

Piatti et al. (1995) derived a very similar gradient of $-0.07 \text{ dex kpc}^{-1}$ from a large and homogeneous sample of abundances from DDO photometry for 63 clusters, primarily younger than 1 Gyr. Unlike previous studies, however, they found clusters to exhibit a steep abundance gradient perpendicular to the galactic plane, of $-0.34 \text{ dex kpc}^{-1}$. Based on the calculation of orbits for 19 clusters with proper motion as well as radial velocity data, they concluded that the present-day gradients have not been modified significantly by cluster orbital motion.

The first appreciable sample of abundances based on spectroscopy came with Friel and Janes (1993), who used low-resolution spectroscopy of cluster red giants, concentrating on the oldest and most distant objects known at the time, and supplemented by several well-studied clusters from the literature, to derive a gradient of $-0.09 \pm 0.02 \text{ dex kpc}^{-1}$. A later paper (Friel et al. 2002) expanded this sample to 39 clusters with a new abundance calibration to find a gradient of $-0.06 \pm 0.01 \text{ dex kpc}^{-1}$ over the range of 7–16 kpc in galactocentric distance (assuming a solar distance of 8.5 kpc). They found no vertical gradient in abundance, once correcting for the selection effect that makes more distant clusters in the outer disk (which are also more metal-poor) more likely to be found at large distances from the plane.

It was common in these, and many other studies of the disk abundance gradient, to fit linear relations to characterize the dependence of abundance with distance as the simplest assumed form, although a number of authors (Friel 1995) commented on the possibility of more complex distributions. Twarog et al. (1997) came to a very different conclusion, using a sample of 76 clusters spanning a wide range of ages, and placed on a common metallicity scale by combining DDO photometry and spectroscopic results from Friel and Janes (1993). They found that the metallicity distribution was best described by two distinct zones in galactocentric radius, separated at 10 kpc, on a scale where the sun is at 8.5 kpc. Between 6.5 and 10 kpc, the cluster metallicity is roughly solar, with a dispersion of only 0.1 dex. Outside 10 kpc, the mean metallicity drops by a factor of 2, to $[\text{Fe}/\text{H}] = -0.3$, and remains constant to the most distant cluster, at ~ 16 kpc. They suggested that the discontinuity is a reflection of the edge of the initial galactic disk, and that the initial offset in $[\text{Fe}/\text{H}]$ created by different histories of chemical evolution on either side of the break has been preserved to the present day in the cluster population.

Since then, the discovery and study of several extremely distant clusters has allowed the gradient to be traced beyond galactocentric distances of 20 kpc and provides insight on the behavior in the very outermost disk. The most distant of these clusters, Berkeley 29, at $R_{gc} = 22$ kpc, has a metallicity $[\text{Fe}/\text{H}]$ of only -0.3 to -0.5 (Yong et al. 2005; Carraro et al. 2004), far more metal-rich than expected if the radial gradient slope of $-0.06 \text{ dex kpc}^{-1}$ continued to



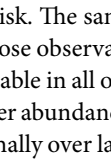
■ Fig. 7-9

Radial abundance gradient from high-resolution spectroscopic studies. Open circles from a variety of sources in the literature, with values as collected by Magrini et al. (2009) supplemented by Friel (2006), filled squares from Sestito et al. (2008), filled circles from Friel et al. (2010), filled triangles from Yong et al. (2005). Clusters in common to samples are connected by dotted lines. Error bars indicate uncertainties in mean cluster abundances as cited by the authors

these distances. As more clusters were added, data showed clearly that the disk radial abundance gradient appears to flatten out beyond about $\sim 10\text{--}12$ kpc, reaching a plateau of $[\text{Fe}/\text{H}] \sim -0.3$ to -0.5 (Yong et al. 2005; Carraro et al. 2007).

Although one must be cautious in merging results from different studies, in the past few years, several of the larger programs aimed at obtaining samples observed and analyzed in a homogeneous and uniform fashion are beginning to yield results. **►** Figure 7-9 combines determinations from a variety of high-resolution studies now available, with the larger samples identified (filled squares from Sestito et al. (2008), filled circles are Friel et al. (2010), and filled triangles are Yong et al. (2005)). Error bars give the error in the mean cluster abundance as cited in the literature; these usually reflect only internal errors. Abundances for clusters in common to more than one study are joined by dotted lines, indicating the magnitude of possible systematic effects between analyses. Although the scatter at any galactocentric radius is appreciable, if one concentrates on the filled points as representative of the more homogeneous samples, there appears to be a fairly steep gradient up to $R_{gc} \sim 10\text{--}13$ kpc, while beyond this distance and out to the most distant cluster at 22 kpc, the slope is consistent with zero. The average metallicity beyond ~ 13 kpc is ~ -0.3 dex.

The nature of the break in behavior of the abundance distribution, and its exact location, is neither well characterized nor understood. Again, emphasizing the importance of large, homogeneous samples, Jacobson and collaborators (e.g., Jacobson et al. 2009) have investigated cluster abundances in the transition region of $\sim 10\text{--}13$ kpc. Their results show substantial scatter in $[\text{Fe}/\text{H}]$ values for clusters in this region, with values ranging from 0.0 to -0.3 , well in excess of observational errors, indicating that the transition is not abrupt as suggested by Twarog et al. (1997).

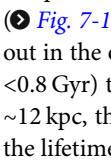
It is worth noting that the value of the gradient in the inner disk is highly dependent on the presence of two very metal-rich clusters NGC 6791 and NGC 6253 at 8 kpc and 7 kpc, respectively. Although the high metallicities of these clusters have been confirmed in numerous studies, and agree within 0.1 dex, it is not clear how representative these clusters are of the inner disk. The sample of clusters shown in  Fig. 7-9 is far from being a complete sample or one whose observational selection biases are well understood.

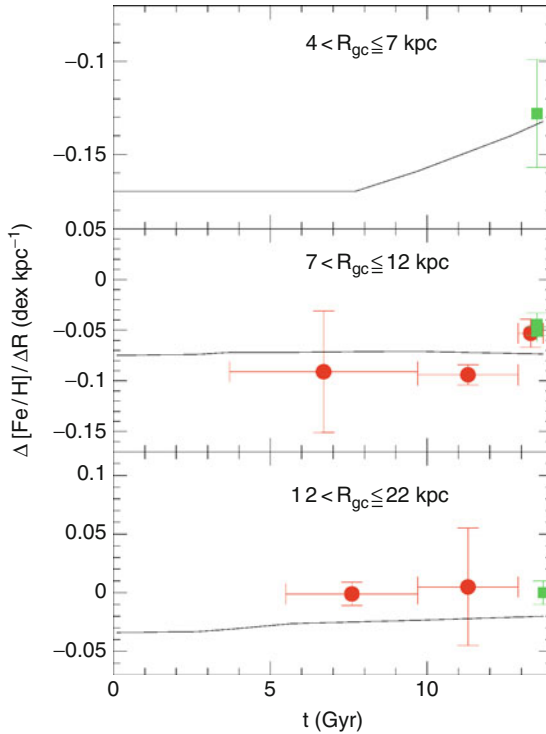
Notable in all of these portrayals of the abundance gradient, however, is the large dispersion in cluster abundances at a given galactocentric radius. The distribution of these samples average azimuthally over large regions in the galactic disk and perhaps it is not unexpected to find large variation in abundance. Are the open clusters telling us about the magnitude of the variation? There is, in fact, evidence for localized and substantial azimuthal variation in abundance in the outer disk. Luck et al. (2006) find, from a sample of almost 200 Cepheids, enhancements of 0.2 dex in $[\text{Fe}/\text{H}]$ over the overall radial gradient, at longitudes of $l \sim 120^\circ$, several kpc exterior to the solar circle. Using young clusters in the inner galaxy, Davies et al. (2009) also find azimuthal variations in abundance that suggest effects due, in this case, to the patchy star formation driven by the central bar of the Milky Way. As the number of confirmed open clusters both in the inner and outer disk continue to increase, they offer the potential to trace the two-dimensional abundance structure of the galactic disk.

4.2 Evolution of the Abundance Gradient with Age

The fact that open clusters can be dated with more certainty than field stars, and that they span all ages in the galactic disk, make them among the best tools with which to study the time evolution of the abundance gradient, a critical constraint to models of galactic chemical evolution.

The earliest studies of the abundance gradient from clusters indicated that the oldest clusters showed the steepest gradients (Janes 1979; Janes et al. 1988). Among their sample of clusters with ages ranging from ~ 700 Myr to 10 Gyr, Friel et al. (2002) found a slight suggestion of a steepening of the gradient with increasing cluster age, but the significance of the result was limited by the restricted distance range for the youngest clusters. The very different distance distributions of clusters of different ages, coupled with the possibility of the gradient changing slope in different parts of the disk complicate the interpretation. The fact that older clusters are found preferentially in the outer disk, and not found at all inside about 6–7 kpc, means that only younger clusters can probe the inner regions of the disk. The discovery and study of clusters to distances of ~ 20 kpc in the disk, however, extends the range over which comparisons can be made.

Magrini et al. (2009) use a compilation of the most recent high-resolution abundance studies of open clusters to investigate the time evolution of the gradient over galactocentric distances of 7–22 kpc. Although by combining disparate samples, systematic differences between studies become a factor, they estimate that these amount to only 0.12 dex in $[\text{Fe}/\text{H}]$. Their sample of 45 clusters span ages of 25 Myr to 11 Gyr when placed on a uniform age scale. With this sample, they find no strong evidence for evolution in the abundance gradient over this period ( Fig. 7-10). In the regions from 7 to 12 kpc, where the gradient is steepest before flattening out in the outer regions, there is slight evidence for the gradient of the youngest clusters (ages < 0.8 Gyr) to be flatter than that determined from the older clusters. In the outer disk, beyond ~ 12 kpc, the cluster abundances are consistent with a zero slope, which does not change over the lifetime of the galaxy.



■ Fig. 7-10

Time evolution of the slope of the abundance gradient, $d[\text{Fe}/\text{H}]/dR$, for three regions of galactocentric distance, from Magrini et al. (2009). *Continuous lines* in each panel are slopes of the gradients predicted by models of galactic chemical evolution. *Filled circles* represent the slopes from open cluster gradients calculated in three age bins: ages <0.8 Gyr, ages between 0.8 and 4 Gyr, and ages >4 Gyr. *Filled squares* represent slopes of gradients from Cepheids calculated for each radial region



4.3 Elemental Abundance Ratios

While overall metallicity and Fe abundances are useful in tracing the general trends of chemical enrichment, they do not provide much information on the details of star formation and nucleosynthesis. Those are revealed in the behavior of elements such as oxygen, the α -elements, r- and s-process elements, and their specific abundance patterns. Although there were a few pioneering efforts to obtain high-resolution spectroscopy from which elemental abundances could be determined, it was not until the mid-1990s and especially the 2000s when studies began to yield appreciable numbers of clusters for which elements other than Fe or the lightest elements (such as Li or C) were determined. Several groups have led in these efforts: the Bologna Open Cluster Chemical Evolution project (Bragaglia and Tosi 2006), Randich and collaborators (e.g., Sestito et al. 2008), Carraro and collaborators (e.g., Carraro et al. 2007), Jacobson, Friel, and collaborators (Jacobson et al. 2009; Friel et al. 2010), and Yong, Carney, and collaborators (Yong et al. 2005).

Of particular interest is the behavior of oxygen and the α -elements Mg, Si, Ca, and Ti, which are formed through stellar nucleosynthetic processes in massive stars, and, as a result, are

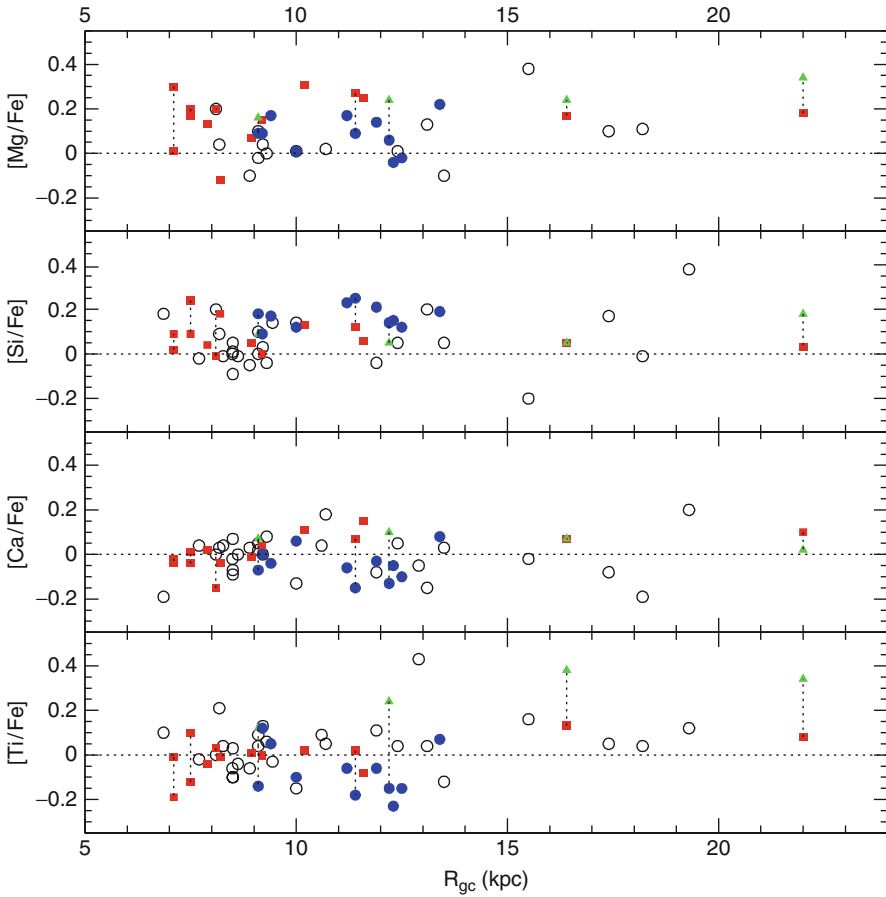
quickly recycled to the interstellar medium through mass loss and Type II supernovae explosions. Elevated abundances of these elements, relative to the solar ratios, point to episodes of rapid star formation and distinctive star formation histories.

Initial studies of the abundances of the outermost disk clusters indicated just this possibility. Yong et al. (2005) and Carraro et al. (2004) found that the clusters that defined the plateau of $[\text{Fe}/\text{H}] \sim -0.3$ in the outer disk also showed elevated values of those elements formed in massive stars, indicating genesis associated with a period of rapid star formation. The pattern of elemental abundances found in these most distant clusters did not match those of either the thin or thick disk, nor the halo. Based on their peculiar abundances, and in particular the enhanced oxygen, α -elements, and the r-process element europium, Yong et al. (2005) suggested that these outer disk open cluster abundance ratios are consistent with the outer disk being formed via a merger event or series of events.

However, further work on these and additional open clusters in the outer disk now suggests abundance ratios that are consistent with scaled solar values, similar to those found in clusters in the solar neighborhood (Carraro et al. 2007; Sestito et al. 2008). The nature of the most distant disk clusters is still a puzzle, but it seems less likely now that they have a distinctly different formation history, at least as revealed in the signature of their abundances.  Figure 7-11 collects results for α -elements from high-resolution studies in the literature, plotting the radial distribution of $[\text{X}/\text{Fe}]$ for Mg, Si, Ca, and Ti. Elemental ratios show no strong dependence on radius, although there are indications that not all α -elements scale in the same way. Mg and Si are often higher than Ca and Ti, but the differences are small and consistent with observational uncertainties. In  Fig. 7-11, clusters in common to different samples are connected by dotted lines and demonstrate the importance of understanding and taking into consideration systematic differences between studies, particularly for those outer disk clusters that indicated enhanced abundances. Overall, the cluster abundances follow the trends of $[\text{Fe}/\text{H}]$ that are seen in field star studies, and fall within the dispersions of the thin disk field star population (see P. E. Nissen, this volume).

There also appear to be no pronounced trends of abundance ratio with age. The oldest clusters, Be 17, NGC 6791, and Cr 261, at ages of 8–10 Gyr, for example, have solar abundance ratios. In particular, their abundances of oxygen and the α -elements, which might have shown some enhancement due to rapid star formation in the early disk, are, instead, consistent with solar values. Over the full range of cluster ages, the α -elements generally show solar ratios as well.

Several elements do show slight enhancements over solar. Na and Al, for example, are often found to be enhanced in open clusters. These elements are produced in burning cycles in intermediate mass stars, in particular the Ne-Na and Mg-Al cycles that operate at high temperatures in shell burning in evolved stars. Enhanced abundances of these elements may point to contributions from winds of intermediate mass stars polluting the interstellar medium with the products of these advanced burning stages. It is also possible that these elements are enhanced through mixing to the stellar surfaces of products of internal processes in evolved stars, since many of the abundance determinations rest on analysis of the brighter cluster giants. As discussed earlier, the limited observations of stars of differing evolutionary state within a cluster do not yet provide a clear picture of whether internal processes are operating. A more prosaic explanation is that the Na enhancement, in particular, may be due to difficulties in analysis, and the lack of correction for non-LTE effects in the abundance analysis of evolved giant stars, which form the basis for most cluster abundance studies. Until these analysis issues are understood fully, the interpretation of the enhanced Na abundances remains unclear.



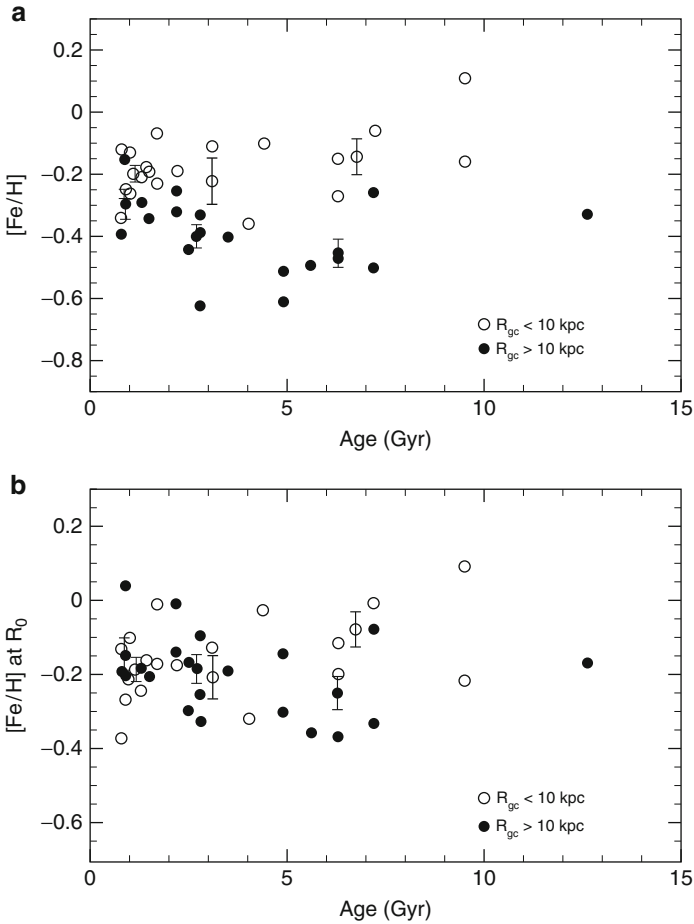
■ Fig. 7-11

Distribution of $[X/Fe]$ for α -elements Mg, Si, Ca, and Ti with galactocentric radius. Open circles are from a variety of sources in the literature, with values as collected by Magrini et al. (2009) supplemented by Friel (2006), filled squares are from Sestito et al. (2008), filled circles are from Friel et al. (2010), filled triangles from Yong et al. (2005). Clusters in common to samples are connected by dotted lines

4.4 Age–Metallicity Relationship

The evolution of the metallicity of the disk is a fundamental observational constraint for theories of galactic chemical evolution. From the first small and local samples, open clusters have posed a challenge to the idea that the galactic disk is gradually enriched over time by the products of successive generations of star formation. The existence of old, solar metallicity clusters such as NGC 188 indicated that the disk had to have been enriched early. As samples have increased, the conclusion has not changed.

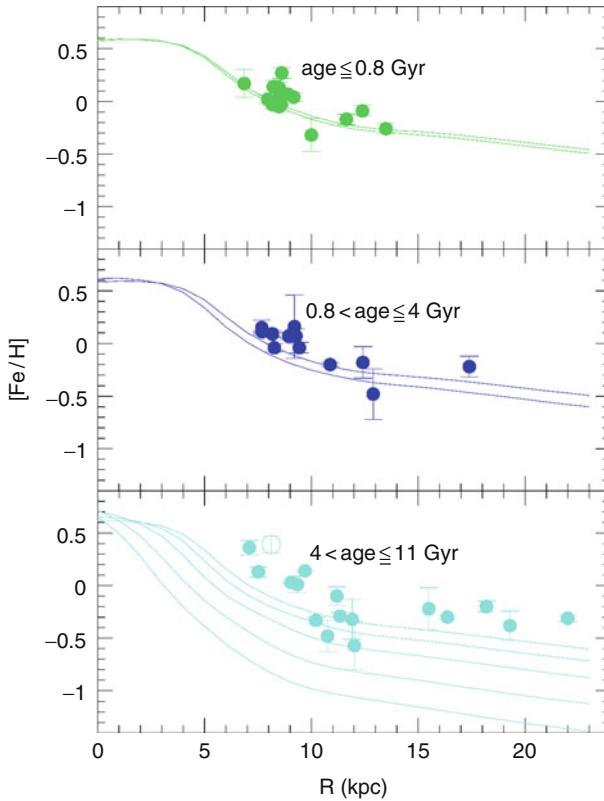
The homogeneous samples of cluster metallicities from Friel and Janes (1993) and Friel et al. (2002) showed no dependence of metallicity on cluster age ranging from Hyades-age to the



■ Fig. 7-12

Open cluster metallicities versus age, from Friel et al. (2002). Clusters are on a uniform relative age scale established in Janes and Phelps (1994). Clusters are distinguished by their position in galactocentric radius; those with $R_{gc} < 10$ kpc are shown as *open circles*; clusters with $R_{gc} > 10$ kpc are shown as *filled circles*. Points with error bars indicate the mean metallicity of clusters in three age ranges: less than 2 Gyr, between 2 and 4 Gyr, and greater than 4 Gyr. Panel (a) shows metallicities as measured. Panel (b) plots metallicities that have been corrected for a radial abundance gradient, assuming a uniform slope of -0.06 dex kpc^{-1}

10 Gyr cluster Be 17 (● Fig. 7-12). Similar results were derived by Carraro and Chiosi (1994) and Salaris et al. (2004), who were careful to place both the cluster metallicities and ages on common scales. Most recently, the compilation of high-resolution results by Magrini et al. (2009), shows that abundances of clusters of different age groups vary very little, if at all, compared to the expectation for enrichment predicted by the chemical evolution models (● Fig. 7-13). While models are able to predict abundances of younger clusters, they fail by a significant margin to predict the elevated abundances of clusters older than a few Gyr, as the lower panel in ● Fig. 7-13 demonstrates.



■ Fig. 7-13

Abundance gradient for open clusters compared to theoretical models from Magrini et al. (2009). Panels show abundances for different age ranges. Solid lines show the predictions for models in each age range: top panel for the present day and 1 Gyr ago, middle panel for 1 and 4 Gyr ago, and lower panel for 4, 6, 8, 10, and 11 Gyr ago

The lack of correlation between cluster age and metallicity is observed at all positions in the disk that are sampled by open clusters, once the overall effect of the radial abundance gradient is taken into account (● Fig. 7-12). Apparently, over the entire age of the disk, at any position in the disk, the oldest clusters form with compositions as enriched as those of much younger objects. The case of NGC 6791 stands as a prime example, with an age of ~ 8 Gyr and a metallicity from high-resolution analyses of $[\text{Fe}/\text{H}] = +0.4$ dex, or twice that of the 700 Myr old Hyades. NGC 6791 may be anomalous in that its orbit indicates a perigalacticon of only 3 kpc (Bedin et al. 2006), so it may have formed in a region of the galaxy much more enriched than its present-day position at a galactocentric distance of 8 kpc. But even discounting NGC 6791 as highly unusual, the roughly solar metallicity of NGC 188, at 6–7 Gyr, and the relatively high metallicity of Be 17, with $[\text{Fe}/\text{H}] = -0.15$ at an age of 10 Gyr, provide other examples that run counter to expectation. It is clear that where a cluster is formed is much more of a factor in determining its metallicity than when it formed.

The distribution of cluster abundances both with respect to age and galactic location offers another important constraint for models of chemical evolution. As [Figs. 7-9](#) and [7-12](#) demonstrate, the open clusters show a rather large dispersion in abundance at a given galactocentric position or at a given age, reflecting a range of abundance of a factor of 2 or more.

4.5 Comparison to the Disk Field Populations

The chemical history of the disk as reflected in the open clusters is, in its general characteristics, very similar to that found in other disk populations. Studies of the radial abundance gradient based on other tracers such as OB stars, H II regions, planetary nebulae, or Cepheids, show dependencies that are consistent with open clusters, recognizing that they track populations in general much younger than the open clusters. Recent work on Cepheids has shown abundances in the outer disk that follow the same flattening in the gradient as revealed in the older open clusters (Yong et al. 2005; Andrievsky et al. 2004). Extensive studies of field stars by Edvardsson et al. (1993) and Nordstrom et al. (2004) show radial gradients in the solar neighborhood that are consistent with those of the open clusters of comparable age.

Considering the age–metallicity relationship, comparison to the field star results is less straightforward to make. While there are many efforts at establishing the relationship between age and metallicity for field stars in the solar neighborhood, debate continues over the shape of the mean relationship, and in particular, the impact of selection effects, biases, or uncertainties that would distort or exclude classes of stars at the extremes of the distributions (see Feltzing et al. 2001 for a review). Although studies agree that there is a large dispersion about any mean relationship, debate continues on whether the mean metallicity shows very little variation with age (as demonstrated in the Nordstrom et al. 2004 sample), or continues to decrease with increasing age (Reddy et al. 2003). Given the large and real scatter in metallicity at all ages in the field star distribution, the lack of dependence of open cluster metallicities with age can be accommodated within any of the extant large field star studies, but is in better overall agreement with the results from Nordstrom’s Geneva–Copenhagen survey.

4.6 Comparison to Theoretical Models

There is now an abundance of models of the chemical evolution of the Milky Way, all of which have had success at reproducing many of the observational constraints provided by the present-day distribution of gas and star density, the star formation rate, the overall run of abundance gradients from early type stars and young disk objects, and the present-day mass function. Although models are increasingly complete, they are not uniquely constrained by the available observational data, and the basic observed relationships most frequently adopted can be reproduced by quite different combinations of model parameters. In all of these models, the primary free parameters are the star formation rate and its dependence on gas and total density, the form of the initial mass function, and the extent, rate, and composition of the gas flows into and out of the galaxy. While the radial abundance gradient, for example, is predicted by any models with a star formation rate that varies with the gas or total mass, its slope is strongly modified by infall.

The open clusters present basic observational behavior that confronts any model of galactic chemical evolution: The decrease in metallicity with increasing Galactocentric radius through

the solar neighborhood to a plateau in the very outer disk, with no strong evidence of any evolution in time, the lack of any correlation between cluster age and metallicity, an appreciable scatter in metallicity at any age and position in the disk, and abundance ratios that are roughly solar and constant with age and location in the disk. While some of these observations add to the evidence from field star studies or other stellar populations, the open clusters provide several constraints that are unique.

Because of their capability to probe a range of ages, particularly the early stages of disk evolution, the open clusters provide the best candidates to understand the time evolution of abundance gradients. Current models of chemical evolution can produce a wide array of predictions of the time evolution of the gradient, depending on the relative importance of the efficiency of star formation and enrichment processes and the nature and amount of infalling material at various locations in the galactic disk (Tosi 1988). For models that predict gradients that steepen with time (e.g., Chiappini et al. 2001), the outer disk is pre-enriched by the previous evolution of the halo, and because of the relatively low star formation, remains relatively little enriched, while the inner disk undergoes more star formation and enrichment over time, resulting in a steeper gradient at later times. For models that show a flattening of the gradient with time (e.g., Hou et al. 2000), the inner disk undergoes rapid star formation which elevates its metallicity initially, while the outer disk undergoes slower enrichment, gradually building up the metallicity in the outer regions relative to the inner, and flattening the gradient. Varying rates of infall of material and levels of pre-enrichment result in different forms and rates of time evolution of the gradient.

The fact that the open clusters show only very slight evidence for the present-day gradient to be slightly flatter than at earlier times, with no strong evolution, coupled with the apparent flattening of the gradient in the outer regions of the galactic disk, favors models that utilize an inside-out formation of the disk. The infall of gas, which is radially dependent, produces a rapid collapse in the inner regions ($R_{gc} < 12$ kpc), with a uniform accretion in the outer regions. Coupled with more efficient star formation in the inner regions, a steeper gradient is established early, while in the outer disk, the low metallicity of the infalling gas and the low star formation rate contribute to maintain a flat and slowly evolving gradient. However, as Magrini et al. (2009) find, to reproduce the completely flat gradient in the outer disk, additional uniform accretion of material is needed, which would result in more star formation and enrichment than is observed. They suggest that the outer plateau could be the result of a past merger which provided pre-enriched material without strongly affecting the star formation rate.

A recent complication to the interpretation of all of these models, however, comes with consideration of the potential effect of radial migration of stars and stellar systems within the disk. Simulations by Roskar et al. (2008) show that stars migrate across significant galactocentric distances due to resonant scattering with transient spiral arms, while preserving their initial circular orbits. Although the models are interpreted in terms of individual stars, the simulations suggest that radial migration mechanisms are capable of affecting stellar clusters as well. As they note, this effect may explain the large scatter and weak correlation seen in the field star and open cluster age–metallicity relationship, and has implications for the shape and evolution of the radial abundance gradient. Radial migration causes more mixing in older populations, resulting in the appearance of flatter gradients at earlier times, but would also dilute the effect of initial gradients. Most importantly, the phenomenon decouples the currently observed properties of stellar populations from their place and conditions of birth and would have a significant impact on the interpretation of cluster properties in a galactic context.

5 Clusters in the Context of Galaxy Formation and Evolution

Our picture of galaxy formation and evolution has matured enormously in the past few decades. Two alternative views continue to offer the principal frameworks in which to consider observational constraints: that of overall halo collapse with subsequent disk formation and continued collapse under self gravity and that of hierarchical accretion and ongoing merger of satellite galaxies within a cold dark matter dominated cosmology (Eggen et al. 1962; Majewski 1993; Freeman and Bland-Hawthorn 2002). There is now ample direct observational evidence of mergers within the Milky Way galaxy, and the fact that they leave distinct dynamical and chemical traces of past events has become a primary tool for investigating and even reconstructing the detailed history of galaxy formation and evolution.

In the framework of the gradual collapse, followed by the self-regulated chemical and dynamical evolution of the gaseous galactic disk, the open clusters are formed from the very earliest stages, as evidenced by the existence of clusters such as NGC 6791 and Be 17. As clusters continue to form within the process of overall star formation, they reflect gradients in disk structural or chemical properties. Clusters are also disrupted by external influences within the disk, such as interactions with molecular clouds, transient spiral arms, and the general tidal field. This interplay between cluster formation and destruction shapes the population now seen, potentially obscuring much of the original distributions that could have preserved the processes of disk formation and evolution.

Nevertheless, there are clues. The trend of decreasing abundance with galactocentric distance, but lack of vertical abundance gradient among the clusters, and the lack of dependence on cluster age tells us about the relative importance of star formation and infall in the disk. The lack of any correlation of cluster age with metallicity and the fact that, at any position in the disk, the oldest clusters are forming with compositions as enriched as those of much younger clusters suggests substantial inhomogeneity in the overall composition of the interstellar material.

If the inside-out model of galaxy formation applies, the open clusters should serve as valuable indicators of the growth of the galactic disk, particularly in the outer regions of the galaxy. One might expect to see the median age of clusters decrease as one moves outward in the galactic disk, pointing to more recent star formation, with young clusters at the very edge of the disk. Indeed, the very oldest clusters known are found inside about 10–11 kpc from the galactic center. However, the open clusters do not show a strong age dependence with R_{gc} and instead the outer disk clusters are generally older than average (☉ Fig. 7-8). This fact is usually interpreted as being due to their ability to survive in these regions and not due to the lack of young clusters. How much of this effect is due to incompleteness in the cluster sample, however, remains to be seen.

In fact, there is evidence for molecular gas and stellar clusters being formed in the far outer regions of the Galaxy. A number of embedded clusters have been found at $R_{gc} \geq 13$ kpc out to 20 kpc. The molecular clouds containing clusters in these regions of low gas density resemble molecular clouds in the inner galaxy, suggesting that processes that lead to the formation of stellar clusters are very similar throughout the galactic disk (Yun et al. 2009). The existence of ongoing cluster formation in the very outer reaches of the disk supports the notion of the gradually growing disk.

The discovery of several open clusters at galactocentric distances of ~ 20 kpc (Be 29 and Saurer 1), along with the observations on the abundance patterns in the outer disk, has led to the idea that distant open clusters may be tracing merger events (Yong et al. 2005). This is not

an unreasonable line of reasoning to pursue. Dwarf galaxies contain star clusters of a variety of ages and the Sagittarius dwarf spheroidal, currently merging with the Milky Way, has at least seven globular clusters associated with it. These clusters are now distributed along the orbital path of the galaxy as it wraps itself around the Milky Way.

If there have been merger events in the galactic disk, might it not be reasonable to expect open clusters to be associated with these mergers, either acquired from the merging galaxies, or formed as a result of the merger event itself? Can any open clusters be associated with known mergers or are there subpopulations of clusters that might reveal past mergers? The current picture is ambiguous.

The discovery in the Sloan Digital Sky Survey of a large overdensity of stars in the region of Canis Major in the outer galactic disk offers the potential to identify star clusters associated with the structure (Yanny et al. 2003; Crane et al. 2003; Conn et al. 2005; Bellazzini et al. 2006). This large stellar structure, known also as the Monoceros Ring, spans almost 100° in galactic longitude, is nearly coplanar with the outer disk, and appears elongated along the tangential direction, with a mean galactocentric radius of $\sim 13\text{--}16$ kpc. The overdensity appears superimposed on the galactic warp, complicating the interpretation. A variety of explanations exist for the structure – simply a manifestation of the galactic warp, an outer spiral arm, a resonance induced by an asymmetric galactic component, a response of the galactic disk to the close passage of a satellite galaxy, or the remnants of a tidally disrupted satellite galaxy. In any case, identifying stellar clusters associated with the stellar structure could help in choosing among these alternatives.

Frinchaboy et al. (2004) identified a number of open and globular clusters, based on both position and limited available radial velocities, that could be associated with the Monoceros Ring and outer disk structure. The cluster planar distribution, argued to be highly significant, suggested to them evidence of an origin related to the interaction of a satellite galaxy with the Milky Way. Later work, along with determination and improvement in radial velocity and distance determinations for the outer disk clusters, suggests that at most only one cluster, Tombaugh 2, has both position and radial velocity that is consistent with an association with the core Canis Major system. However, its properties are also entirely in keeping with those of open clusters at its general galactocentric radius; its metallicity and its elemental abundance pattern do not distinguish it in any way from the overall cluster population (Villanova et al. 2010). Whether other clusters may be associated with the extended stellar stream of a disrupting satellite galaxy remains to be seen.

The two most distant open clusters in the galaxy, Be 29 and Saurer 1, have recently been claimed to be associated with the Sgr dwarf (Carraro and Bensby 2009), based on a combination of their locations and velocities being consistent with models of the trailing stream of the disrupting galaxy. The cluster abundances present a confused picture, however, and do not distinguish the clusters in terms of star formation history that might associate them uniquely with a stellar population other than the general outer disk.

At present, while intriguing, there appears to be no strong evidence that would associate any open clusters with specific merging events in the galactic disk.

Even if no open clusters have yet been definitively associated with the remnants of an accretion event, some of the properties of old clusters are naturally explained in the context of merger scenarios. Evidence for infall to the disk comes in the form of not only accretion events, but high velocity clouds and the energetic recycling of processed material from the disk. Theoretical modeling indicates that these events can initiate star formation at appreciable distances from the plane, and should clusters form, they would preserve the large z motions and

possibly eccentric orbits introduced by the colliding or accreting material. This scenario provides an explanation for the existence of old open clusters at substantial distances from the galactic plane and a mechanism for them to have acquired the orbits that allow them to survive to such substantial ages. Very few open clusters have proper motion and radial velocity measurements that allow the determination of their orbits to test this idea. The handful of clusters with radial velocities that are suggestive of eccentric or unusual orbits would be particularly interesting to investigate.

Cross-References

- [Dynamics of Disks and Warps](#)
- [Galactic Distance Scales](#)
- [Globular Cluster Dynamical Evolution](#)
- [Interstellar PAHs and Dust](#)
- [Mass Distribution and Rotation Curve in the Galaxy](#)
- [Star Counts and the Nature of the Galactic Thick Disk](#)

References

- Andersen, J., Nordstrom, B., & Clausen, J. V. 1990, *ApJL*, 363, 33
- Andrievsky, S. M., Luck, R. E., Martin, P., & Lepine, J. R. D. 2004, *A&A*, 413, 159
- Ascenso, J., Alves, J., & Lago, M. T. V. T. 2009, *A&A*, 495, 147
- Bedin, L. R., Piotto, G., Carraro, G., King, I. R., & Anderson, J. 2006 *A&A*, 460, 27
- Bellazzini, M., Ibata, R., Martin, N., Lewis, G. F., Conn, B., & Irwin, M. J. 2006, *MNRAS*, 366, 865
- Bertelli, G., Bressan, A., & Chiosi, C. 1985, *A&A*, 150, 33
- Bica, E., Dutra, C. M., & Barbuy, B. 2003, *A&A*, 397, 177
- Bonato, C., & Bica, E. 2007, *MNRAS*, 377, 1301
- Bonato, C., & Bica, E. 2008, *A&A*, 485, 81
- Bonato, C., Kerber, L. O., Bica, E., & Santiago, B. X. 2006, *A&A*, 446, 121
- Bouvier J., Kendall, T., Meeus, G., Testi, L., Moraux, E., Stauffer, J. R., James, D., Cuillandre, J.-C., Irwin, J., McCaughrean, M. J., Baraffe, I., & Bertin, E. 2008, *A&A*, 481, 661
- Bragaglia, A., & Tosi, M. 2006, *AJ*, 131, 1544
- Bragaglia, A., Tosi, M., Andreuzzi, G., & Marconi, G. 2006, *MNRAS*, 368, 1971
- Carraro, G., & Bensby, T. 2009, *MNRAS*, 397, L106
- Carraro, G., Bresolin, G., Villanova, S., Matteucci, F., Patat, F., & Romaniello, M. 2004, *AJ*, 128, 1676
- Carraro, G., & Chiosi, C. 1994, *A&A*, 288, 751
- Carraro, G., Geisler, D., Villanova, S., Frinchaboy, P. M., & Majewski, S. R. 2007, *A&A*, 476, 217
- Carraro, G., Janes, K. A., & Eastman, J. D. 2005, *MNRAS*, 364, 179
- Carraro, G., Ng, Y. K., & Portinari, L. 1998, *MNRAS*, 296, 1045
- Chiappini, C., Matteucci, F., & Romano, D. 2001, *ApJ*, 554, 1044
- Conn, B. C., Lewis, G. F., Irwin, M. J., Ibata, R. A., Ferguson, A. M. N., Tanvir, N., & Irwin J. M. 2005, *MNRAS*, 362, 475
- Crane, J. D., Majewski, S. R., Rocha-Pinto, H. J., Frinchaboy, P. M., Skrutskie, M. F., & Law, D. R. 2003, *ApJ*, 594, L119
- Currie, T., Hernandez, J., Irwin, J., Kenyon, S. J., Tokarz, S., Balog, Z., Bragg, A., Berlind, P., & Calkins, M. 2010, *ApJ Supp*, 186, 191
- Daniel, S. A., Latham, D. W., Mathieu, R. D., & Twarog, B. A. 1994, *PASP*, 106, 281
- Davies, B., Origlia, L., Kudritzki, R.-P., Figer, D. F., Rich, R. M., Najarro, F., Negueruela, I., & Clark, J. S. 2009, *ApJ*, 696, 2014
- de la Fuente Marcos, R. 1997, *A&A*, 322, 764
- De Silva, G. M., Sneden, C., Paulson, D. B., Asplund, M., Bland-Hawthorn, J., Bessell, M. S., & Freeman, K. C. 2006, *AJ*, 131, 455
- Dias, W. S., Alessi, B. S., Moitinho, A., & Lepine, J. R. D. 2002, *A&A*, 389, 871
- Edvardsson, B., Andersen, J., Gustafsson, B., Lambert, D. L., Nissen, P. E., & Tomkin, J. 1993, *A&A*, 275, 101
- Eggen, O. J., Lynden-Bell, D., & Sandage, A. R. 1962, *ApJ*, 136, 748

- Feltzing, S., Holmberg, J., & Hurley, J. R. 2001, *A&A*, 377, 911
- Freeman, K., & Bland-Hawthorn, J. 2002, *ARAA*, 40, 487
- Friel, E. D. 1995, *ARAA*, 33, 381
- Friel, E. D. 2006, in *Chemical Abundances and Mixing in the Milky Way and its Satellites*, Castiglione della Pescaia, Italy. ESO Astrophysics Symposia (Springer-Verlag), 3
- Friel, E. D., Jacobson, H. R., & Pilachowski, C. A. 2010, *AJ*, 139, 1942
- Friel, E. D., & Janes, K. A. 1993, *A&A*, 267, 75
- Friel, E. D., Janes, K. A., Tavaréz, M., Scott, J., Katsanis, R., Lotz, J., Hong, L., & Miller, N. 2002, *AJ*, 124, 2693
- Frinchaboy, P. M., Majewski, S. R., Crane, J. D., Reid, I. N., Rocha-Pinto, H. J., Phelps, R. L., Patterson, R. J., & Munoz, R. R. 2004, *ApJL*, 602, 21
- Froebrich, D., Meusinger, H., & Scholz, A. 2008, *MNRAS*, 390, 1598
- Froebrich, D., Scholz, A., & Raftery, C. L. 2007, *MNRAS*, 374, 399 (FSR)
- Geller, A. M., Mathieu, R. D., Harris, H. C., & McClure, R. D. 2008, *AJ*, 135, 2264
- Gieles, M., Portegies Zwart, S. F., Baumgardt, H., Athanassoula, E., Lamers, H. J. G. L. M., Sipior, M., & Leenaarts, J. 2006, *MNRAS*, 371, 793
- Gozzoli, E., Tosi, M., Marconi, G., & Bragaglia, A. 1996, *MNRAS*, 283, 66
- Grocholski, A. J., & Sarajedini, A. 2003, *MNRAS*, 345, 1016
- Harayama, Y., Eisenhauer, F., & Martins, F. 2008, *ApJ*, 675, 1319
- Hou, J. L., Prantzos, N., & Boissier, S. 2000, *A&A*, 362, 921
- Hurley, J. R., Pols, O. R., Aarseth, S. J., & Tout, C. A. 2005, *MNRAS*, 363, 293
- Jacobson, H. R., Friel, E. D., & Pilachowski, C. A. 2009, *AJ*, 137, 4753
- Janes, K. A. 1979, *ApJ Suppl*, 39, 135
- Janes, K. A., & Phelps, R. L. 1994, *AJ*, 108, 1773
- Janes, K. A., Tilley, C., & Lynga, G. 1988, *AJ*, 95, 771
- Kalirai, J. S., Fahlman, G. G., Richer, H. B., & Ventura, P. 2003, *AJ*, 126, 1402
- Kalirai, J. S., Hansen, B. M. S., Kelson, D. D., Reitzel, D. B., Rich, R. M., & Richer, H. B. 2008, *ApJ*, 676, 594
- Kaluzny, J., & Udalski, A. 1992, *Acta Astron*, 42, 29
- Kim, S. S., Figer, D. F., Kudritzki, R. P., & Najarro, F. 2006, *ApJ*, 653, 113
- King, I. 1962, *AJ*, 67, 471
- King, I. R. 1966, *AJ*, 71, 64
- Kraus, A. L., & Hillenbrand, L. A. 2007, *AJ*, 134, 2340
- Kroupa, P. 2002, *Science*, 295, 82
- Krusberg, Z. A. C., & Chaboyer, B. 2006, *AJ*, 131, 1565
- Lada, C. J., & Lada, E. A. 2003, *ARAA*, 41, 57
- Lamers, H. J. G. L. M., & Gieles, M. 2006, *A&A*, 455, L17
- Luck, R. E., Kovyukh, V. V., & Andrievsky, S. M. 2006, *AJ*, 132, 902
- Lynga, G. 1982, *A&A*, 109, 213
- Maeder, A., & Meynet, G. 1991, *A&A Suppl*, 89, 451
- Magrini, L., Sestito, P., Randich, S., & Galli, D. 2009, *A&A*, 494, 95
- Majewski, S. R. 1993, *ARAA*, 31, 575
- Mathieu, R. D. 1984, *ApJ*, 284, 643
- Mathieu, R. D. 1985, in *IAU Colloq. 88, Stellar Radial Velocities*, ed. A.G. Davis Philip, & D. W. Latham (Schenectady, NY: L. Davis Press), 249
- Mathieu, R. D., & Geller, A. M. 2009, *Nature*, 462, 1032
- McMillan, S. L. W., Vesperini, E., & Portegies Zwart, S. F. 2007, *ApJL*, 655, 45
- Mercer, E. P., Clemens, D. P., Meade, M. R., et al. 2005, *ApJ*, 635, 560
- Mermilliod, J.-C., Mayor, M., & Udry, S. 2008, *A&A*, 498, 949
- Montgomery, K. A., Marschall, L. A., & Janes, K. A. 1993, *AJ*, 106, 181
- Moraux, E., Bouvier, J., Stauffer, J. R., Barrado y Navascues, D., & Cuillandre, J.-C. 2007, *A&A*, 471, 499
- Moraux, E., Kroupa, P., & Bouvier, J. 2004, *A&A*, 426, 75
- Nordstrom, B., Andersen, J., & Andersen, M. I. 1997, *A&A*, 322, 460
- Nordstrom, B., Mayor, M., Andersen, J., Holmberg, J., Pont, F., Jorgensen, B. R., Olsen, E. H., Udry, S., & Mowlavi, N. 2004, *A&A*, 418, 989
- Oort, J. H. 1958, in *Stellar Populations, Recherche Astronomique, Specola Vaticana*, ed. D. J. K. O'Connell. (Amsterdam: North Holland), 63
- Pasquini, L., Randich, S., Zoccali, M., Hill, V., Charbonnel, C., & Nordstrom, B. 2004, *A&A*, 424, 951
- Phelps, R. L. 1997, *ApJ*, 483, 826
- Phelps, R. L., Janes, K. A., & Montgomery, K. A. 1994, *AJ*, 107, 1079
- Piatti, A. E., Claria, J. J., & Abadi, M. G. 1995, *AJ*, 110, 2813
- Pinsonneault, M. 1997, *ARAA*, 35, 557
- Portegies Zwart, S., Gaburov, E., Chen, H.-C., & Guran, M. Atakan 2007, *MNRAS*, 378, 29
- Portegies Zwart, S. F., Hut, P., McMillan, S. L. W., & Makino, J. 2004, *MNRAS*, 351, 473
- Portegies Zwart, S. F., McMillan, S. L. W., Hut, P., & Makino, J. 2001, *MNRAS*, 321, 199
- Randich, S., Sestito, P., Primas, F., Pallavicini, R., & Pasquini, L. 2006, *A&A*, 450, 557
- Reddy, B. E., Tomkin, J., Lambert, D. L., & Allende Prieto, C. 2003, *MNRAS*, 340, 304

- Roskar, R., Debattista, V. P., Quinn, T. R., Stinson, G. S., & Wadsley, J. 2008, *ApJL*, 684, 79
- Salaris, M., Weiss, A., & Percival, S. M. 2004, *A&A*, 414, 163
- Salpeter, E. 1955, *ApJ*, 121, 161
- Schuler, S. C., King, J. R., & The, L.-S. 2009, *ApJ*, 701, 837
- Schuler, S. C., Hatzes, A. P., King, J. R., Kurster, M., & The, L.-S. 2006, *AJ*, 131, 1057
- Scott, J. E., Friel, E. D., & Janes, K. A. 1995, *AJ*, 109, 1706
- Sestito, P., Bragaglia, A., Randich, S., Pallavicini, R., Andrievsky, S. M., & Korotin, S. A. 2008, *A&A*, 488, 943
- Slesnick, C. L., Hillenbrand, L. A., & Massey, P. 2002, *ApJ*, 576, 880
- Smiljanic, R., Gauderon, R., North, P., Barbuy, B., Charbonnel, C., & Mowlavi, N. 2009, *A&A*, 502, 267
- Spitzer, L. 1958, *ApJ*, 127, 17
- Spitzer, L., & Harm, R. 1958, *ApJ*, 127, 544
- Stolte, A., Brandner, W., Brandl, B., & Zinnecker, H. 2006, *ApJ*, 132, 253
- Terlevich, E. 1987, *MNRAS*, 224, 193
- Tosi, M. 1988, *A&A*, 197, 33
- Twarog, B. A., Ashman, K. M., & Anthony-Twarog, B. J. 1997, *AJ*, 114, 2556
- VandenBerg, D. A., & Stetson, P. B. 2004, *PASP*, 116, 997
- van den Bergh, S. 1958, *Z. Astrophys.*, 46, 176
- van den Bergh, S. 2006, *AJ*, 131, 1559
- van den Bergh, S., & McClure, R. D. 1980, *A&A*, 80, 360
- Villanova, S., Randich, S., Geisler, D., Carraro, G., & Costa, E. 2010, *A&A*, 509, 102
- von Hippel, T., Jefferys, W. H., Scott, J., Stein, N., Winget, D. E., DeGennaro, S., Dam, A., & Jeffery, E. 2006, *ApJ*, 645, 1436
- Yanny, B., Newberg, H. J., Grebel, E. K., Kent, S., Odenkirchen, M., Rockosi, C. M., Schlegel, D., Subbarao, M., Brinkmann, J., Fukugita, M., Ivezić, Z., Lamb, D. Q., Schneider, D. P., & York, D. G. 2003, *ApJ*, 588, 824
- Yong, D., Carney, B. W., & De Almeida, M. L. T. 2005, *AJ*, 130, 597
- Yong, D., Carney, B. W., De Almeida, M. L. T., & Pohl, B. L. 2006, *AJ*, 131, 2256
- Yun, J. L., Davide, E., Palmeirim, P. M., Gomes, J. I., & Martins, A. M. 2009, *A&A*, 500, 833

8 Star Counts and Nature of the Galactic Thick Disk

Yuzuru Yoshii

Institute of Astronomy, School of Science, University of Tokyo,
Tokyo, Japan

1	<i>Introduction: Historical Overview</i>	395
2	<i>The Star Count Galaxy Model</i>	404
2.1	Fundamental Equation	405
2.2	Input Data	407
2.2.1	Solar Metallicity Inputs	407
2.2.2	Lower Metallicity Inputs	409
2.3	Functional Forms	410
2.3.1	Density Function	410
2.3.2	Metallicity Function	412
2.3.3	Model Parameters	412
2.4	Star Count Observations	413
3	<i>Structural Constraints from Star Counts</i>	415
3.1	Estimates of Structural Parameters	415
3.2	Implications of Recent Estimates	416
3.3	Vertical Scale Height of Thick Disk	417
3.4	Radial Scale Length of Thick Disk	418
3.5	Systematics in the Results	419
3.5.1	PPA or MFA	419
3.5.2	Binary Effect	420
4	<i>Interpretation of the Hess Diagram</i>	420
5	<i>Other Constraints on the Thick Disk</i>	426
5.1	Kinematical Constraints	427
5.2	Metallicity Distribution Constraints	429
5.3	Age Constraints	430
5.4	Elemental Abundance Constraints	430
6	<i>Formation and Evolution of Thick Disk</i>	431
6.1	Summary of Observational Constraints of the Thick Disk	431
6.2	Possible Scenarios	433
6.3	Origin of Double Exponential Stellar Disk	434
6.3.1	Vertical Exponentiality	435

6.3.2	Radial Exponentiality	436
6.4	Transition from Halo to Thick Disk	439
7	<i>Future Directions</i>	441
	<i>Acknowledgments</i>	442
	<i>References</i>	443

Abstract: Modern star counts at high Galactic latitudes played a major role in revealing the existence of a thick disk as the third stellar component of the Milky Way Galaxy in addition to the old thin disk and halo. A number of star count observations and models showed that the thick disk is represented well by a double exponential density law in the vertical and radial directions. The thick-disk structural parameters determined to date from star count analysis are reviewed, and their limitations are described in terms of the correlation among the derived parameters. The recent preference for $h_Z \sim 0.7$ kpc for the scale height of the thick disk, associated with $f_{\text{thick}} \sim 0.1$ for its normalization relative to the thin disk, is likely a consequence of the recent popularity of the flattened inner halo with an axial ratio of $q \sim 0.6$ prescribed in star-count modeling. This value of h_Z for the thick disk is supported by the kinematic constraint of ~ 40 km s $^{-1}$ for the measured vertical velocity dispersion of candidate thick-disk stars more than 1 kpc from the disk plane. Furthermore, star counts in multiple directions and from all-sky near-infrared surveys have arrived at a convergent result, indicating that the thick disk has a scale length $h_R \sim 3.5$ kpc and has a greater radial extension compared to the thin disk, with $h_R \sim 2.5$ kpc. Other constraints have arisen from high-resolution spectroscopic observations of the kinematics, chemical abundances, and ages of candidate thick-disk stars, confirming the rotational lag of ~ 40 km s $^{-1}$ as well as the vertical gradients of the mean rotation and velocity dispersions in three directions, the constant ratio of alpha to the iron abundances $[\alpha/\text{Fe}]$ of $\sim +0.4$ dex up to $[\text{Fe}/\text{H}] \sim -0.4$ dex, a large scatter of metallicity around the mean $[\text{Fe}/\text{H}] \sim -0.8$ dex with little or no spatial gradient, and a fairly old thick-disk age of ~ 10 Gyr. The star counts and other constraints together indicate dissipational contraction and spin-up of an extended disk-like gas component or early gas-rich mergers in which thick-disk stars formed in situ with more rapid chemical enrichment than in the thin disk. Successful scenarios of thick-disk formation and evolution must address all these constraints and furthermore involve a self-regulating mechanism that produces a universal double exponential stellar structure for both thin and thick disks in spiral galaxies.

Keywords: Galaxies: spiral, Galaxy: disk, Galaxy: evolution, Galaxy: formation, Galaxy: fundamental parameters, Galaxy: kinematics and dynamics, Galaxy: structure

1 Introduction: Historical Overview

The thick disk has been a controversial, intermediate field star component between the two major components, the disk and the halo, which are clearly distinct from each other; its reality has long been debated. These components accommodate different populations of stars grouped according to similar spatial and physical characteristics.

Except for young stars currently forming near the midplane of the disk, most stars now observed in either the disk or the halo are less massive and are long-lived enough to have survived throughout the age of the Galaxy. Stars, once formed, orbit adiabatically around the galactic center, which might have grown in mass by subsequent central concentration. At the same time, the chemical abundances in a star's atmosphere change little unless the rare event of efficient chemical pollution of the stellar surface happens to occur through encounters with interstellar clouds. Therefore, the kinematical and chemical properties of old stars retain a fossil record of ancient processes occurring in the galaxy to which they belong. In other words, apparent differences in stellar populations, when arranged in order by age dating, reflect sequential

changes in the kinematical and chemical properties with respect to time, that is, the evolution of the Galaxy.

The concept of stellar populations was introduced by Baade (1944), who for the first time resolved individual stars on photographic plates of the Andromeda Galaxy (M31) and classified them into Populations I and II. By definition, Population I represents disk stars that follow an exponential surface brightness profile in the vertical direction away from the midplane of the disk as well as in the radial direction away from the galactic center, and Population II represents halo stars that follow a more centrally concentrated brightness profile. These two populations together roughly describe spiral galaxies in general, including our Galaxy.

In the disk and the halo, if each is relaxed in dynamical equilibrium, the spatial distribution of stars is related to their kinematics according to the collisionless Boltzmann equation associated with the Poisson equation at different degrees of dark matter contribution in the region that encloses the disk or halo stars investigated. In particular, the shape of the spatial distribution of stars is determined by the relative amounts of random and systematic motion in the total kinematic energy to which the gravitational potential energy is equated. Thus, whereas halo stars show an extended near-spherical spatial distribution with dominant nonordered motion or random motion, disk stars exhibit a highly flattened spatial distribution with dominant ordered motion or systematic rotation. Because the components are never isolated from each other but are mutually related through gravitational attraction, stars in the transition region between disk and halo are expected to exhibit characteristics intermediate between those of the disk and the halo.

Given that any smooth transition is in principle described by the summation of many intermediate components having different weights, we ask what minimum number of them is required to represent the stars in this region of transition. The usual practice in astronomy would be to first propose a single component to represent all of the stars in the intermediate region for simplicity, and later explore the possibility of further division of this single component into several components.

Therefore, in addition to examining the spatial distribution of the intermediate component or thick disk, it is extremely important to examine whether a continuous trend or spatial gradient exists in the kinematics, chemical abundances, and ages of thick disk stars. If a continuous trend in these quantities is assumed, the thick disk might be considered to have formed between the extended halo and the thin disk. Alternatively, if a scatter is observed with no continuous trend, the thick disk could be considered the result of nonsequential, external impacts to the thin disk. These considerations certainly motivated later efforts at obtaining massive amounts of data on the kinematics, chemical abundances, and ages of stars that presumably belong to the thick disk.

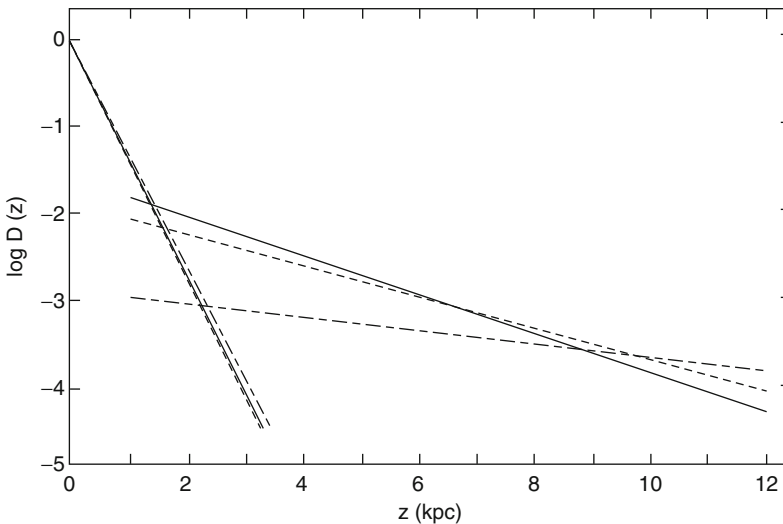
Historically, it has been known since the Vatican Conference held in 1957 that Populations I and II did not account for all of the stars in the Galaxy, but intermediate populations were also required (O'Connell 1958). The density distribution of stars perpendicular to the disk plane was then actively investigated by many authors, for example, using star count data for K giants (Oort 1960) and RR Lyrae variable stars (Plaut 1965), as well as those of main-sequence stars of FG spectral types (Elvius 1965; Uggren 1963). However, the limiting magnitude of their samples was 13 mag, and undoubtedly no intermediate populations would have been evident even in the data for K giants, which reached at most 1.5–2 kpc from the sun, whereas some of the variable stars in the solar neighborhood were known to belong to the intermediate component, judging from their kinematics (Blaauw 1965).

A definite resolution regarding whether intermediate populations would be only a minor fraction of the stars in the Galaxy had to wait until the early 1980s, when modern technology

enabled a major advance in sufficiently accurate photometry of an enormous number of faint field stars in the Galaxy. In particular, the study of star count data consisted of recording the magnitude of every stellar image as faint as 22 mag in a finite area of sky. This was made possible by the advent of automated high-precision measuring machines and high-speed computers. Furthermore, it was recognized that, by obtaining color information in addition to magnitude information, a far more sensitive test of the Galaxy's spatial structure could be made.

Rapid progress in modeling such massive star count data was triggered by timely publication of a seminal paper by Bahcall and Soneira (1980), who constructed a sophisticated Galaxy model with only two components, the disk and the halo, assuming the solar-neighborhood luminosity function (LF) throughout the Galaxy. They emphasized that these two components would suffice to explain all the existing star count data brighter than 22 mag in various directions on the sky. Their model represented disk stars by a double exponential density law with a scale height of 325 pc and a scale length of 4 kpc, and halo stars by the deprojected de Vaucouleurs $r^{1/4}$ law with an effective scale radius of 5 kpc. The number density of halo stars of 0.00125 (= 1/800) of the disk in the solar neighborhood was adopted (Schmidt 1975).

Yoshii (1982) fitted a similar Galaxy model to the Basel star count data in the U , G , and R bands brighter than $G = 19$ mag in an area of 2.61 deg^2 in Kapteyn Selected Area (SA) 57 near the north Galactic pole (NGP) and separated a second component relative to the old disk, assuming possible contamination of subgiants in the sample and an LF that declined steeply toward brighter absolute magnitudes like those of globular clusters (see [▶ Fig. 8-1](#)). The exponential scale height of this component was found to be 2 kpc, and the normalization



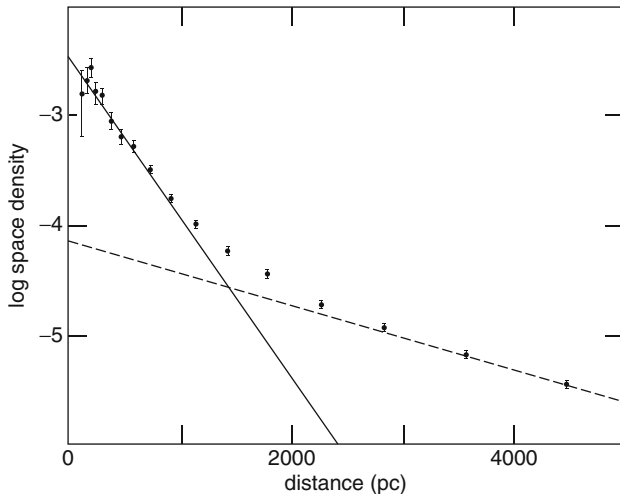
■ Fig. 8-1

Density distributions for the old disk and the second disk component in the vertical direction as a function of distance from the Galactic plane, taken from Fig. 4 of Yoshii (1982). Models are fitted to the Basel star count data toward the north Galactic pole. *Solid and dashed lines* represent models with the globular cluster luminosity function for second-disk stars. *Long-and-short-dashed line* represents the model with the solar-neighborhood luminosity function throughout the Galaxy, but this model was ruled out by Yoshii by comparing it with the observed color distribution

was 0.01–0.02 of the old disk in the solar neighborhood, well beyond Schmidt’s halo normalization by a factor of 10. This component, which Yoshii originally called the halo, was certainly *not* the halo but was identified with the intermediate Population II between the old disk and the halo.

Yoshii’s finding was supported by Gilmore and Reid (1983), who found a significant slope change in the vertical density law at heights of $Z \sim 1\text{--}2$ kpc toward the south Galactic pole (SGP) using the UK Schmidt star count data in the V and I bands brighter than $I = 18$ mag in an area of 18.24 deg^2 (see [► Fig. 8-2](#)). They derived the absolute magnitudes of the stars from their photometric parallaxes, assuming that all the stars are on the main sequence. Direct estimates of stellar distances revealed a change in the brighter part of the LF with increasing vertical distance above the Galactic plane, like those observed in globular clusters near the disk. Then, Gilmore and Reid identified a component having an exponential scale height of 300 pc as the old disk and distinguished a second component having an exponential scale height of 1.45 kpc and normalization of 0.02. They called this second component the thick disk, which is the same terminology used for the disks found in several edge-on S0 galaxies by Burstein (1979). In contrast with the thick disk, the old disk has been called the thin disk.

The need for the intermediate component or thick disk, which was independently derived from Yoshii’s model fit analysis (MFA) and Gilmore and Reid’s photometric parallax analysis (PPA), conflicted with Bahcall and Soneira’s preference for a Galaxy model with only two components, the thin disk and halo. Bahcall and Soneira (1984) claimed that a Galaxy model with a thin disk and a thick disk could not simulate the color distribution of stars fainter than 20 mag. In light of this fact, Gilmore (1984) proposed a Galaxy model with three components, the thin



■ Fig. 8-2

Density distributions for the old disk and the second disk component in the vertical direction as a function of distance from the Galactic plane, taken from Fig. 6 of Gilmore and Reid (1983). Filled circles represent the results for stars with $4 < M_V < 5$ toward the south Galactic pole, based on photometric parallax analysis. Solid and broken lines correspond to the old disk and the second disk, respectively

disk, thick disk, and halo. Bahcall et al. (1985) showed that all the Basel *UGR* star count data brighter than $G = 19$ mag are compatible with Bahcall and Soneira's two-component model and also with Gilmore's three-component model.

To resolve the disagreement regarding this controversial thick-disk component (Bahcall and Soneira 1984; Gilmore 1984), the acquisition of new star count data of high accuracy over a sufficiently large celestial area to minimize the effect of local fluctuations in the spatial distribution of stars was recommended. In particular, star count data in the polar regions are optimal because the line of sight emerges from the thin disk in the shortest distance, the amount of interstellar extinction is the smallest, and any confusion effect in the detection of stellar images is minimal.

The sampled region of space that contributes most to the star count data is located at the distance Z_m at which $D(Z)Z^2$ becomes maximum, where $D(Z)$ is the density law as a function of the distance from the sun to the polar region. The exponential scale height h_z gives $Z_m = 2h_z$, corresponding to a distance modulus of 9 mag for the thin disk, 12 mag for the thick disk, and 15 mag for the halo. With an absolute magnitude of +5 near the turnoff, a complete sample of field stars in a magnitude range of 16–18 mag in the polar regions is expected to account for a large fraction of the thick disk stars. Thus, the frequency distributions of various colors of stars in this magnitude range are very helpful in providing further constraints on the nature of the thick disk. For example, the blue ridge of the $B - V$ color distribution constrains the turnoff color and hence the age of the thick disk, whereas the blue ridge in the $U - B$ color distribution stringently constrains its metallicity owing to the U -sensitive effect of line blanketing in the stellar atmosphere. It is worth noting that the sharpness on the bluer side of the $U - B$ distribution off the ridge measures the level of the spatial gradient of metallicity in the thick disk.

Mainly for these reasons, Stobie and Ishida (1987) obtained the Kiso Schmidt star count data in the U , B , and V bands brighter than $V = 18$ mag in 20 deg^2 in SA 57 near the NGP region. By fitting the models to these data, Yoshii et al. (1987) reinforced the need for a thick disk that makes up 0.02 of the thin disk with a scale height of about 1 kpc and a vertical metallicity gradient as small as $-0.1 \text{ dex kpc}^{-1}$. In fact, Gilmore's three-component Galaxy model was superior to Bahcall and Soneira's two-component model.

The Galactic thick disk was thus discovered and defined by counting stars in the NGP and SGP regions, but a very loose constraint was imposed on its radial structure by off-polar observations, that is, $h_R > 2.5 \text{ kpc}$ even for an exponential scale length of the thin disk (Bahcall and Soneira 1984). Moreover, the scale length of the thick disk was at best assumed to be the same as that of the thin disk. This motivated further acquisition of star count data in various directions. Under the usual assumption of rotational symmetry and north–south symmetry of the Galaxy, a pair of directions toward the center and anticenter with the same latitude $|b|$ in the sectional plane for $l = 0^\circ$ and 180° is ideal for constraining the scale length in the radial direction.

Except for the best-studied polar regions (e.g., Chiu 1980; Gilmore and Reid 1983; Kron 1980; Stobie and Ishida 1987; Tyson and Jarvis 1979), the BVI star count data brighter than $V = 19$ mag over a sufficiently wide area were obtained only in the 22 h field toward the center $(l, b) = (30^\circ, -51^\circ)$ over 17 deg^2 (Gilmore et al. 1985). To supplement these data, the UBV star count data brighter than $V = 18$ mag were obtained in the anticenter direction of SA 54, $(l, b) = (200^\circ, +59^\circ)$ over 16 deg^2 (Yamagata and Yoshii 1992). Although the scale length of the disk was indeed constrained to be 4 kpc, it was not possible to guarantee separate constraints on the scale lengths of the thin and thick disks.

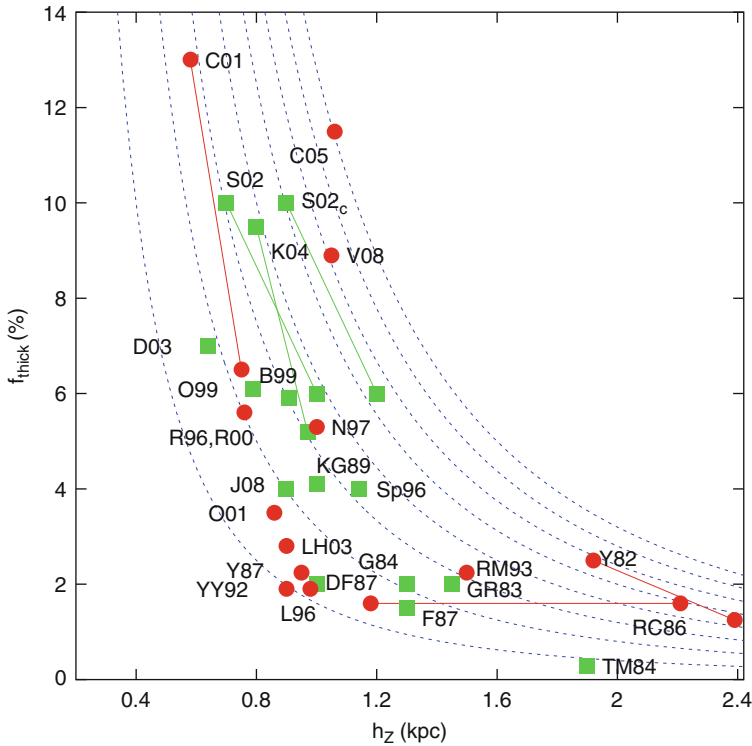
Until the late 1980s, there was no room for doubt about the existence of a thick disk, and the observations and discussions that followed increased not only the complexities of its nature but also the mysteries of its origin. **Table 8-1** summarizes the Galactic structural parameters

Table 8-1
Previous star count results

h_z (pc)	Thick disk			Halo				Method	Tracer	Band	Reference
	h_R (kpc)	f_{thick}	h_z (kpc)	h_R (kpc)	f_{halo}	Density law	r_e (kpc)				
310–325	–	0.0125–0.025	1.92–2.39	–	–	–	–	–	dwarfs, giants	UGR	Yoshii (1982, Y82)
300	–	0.02	1.45	–	0.0020	deV	3.0	0.85	dwarfs	BVI	Gilmore and Reid (1983, GR83)
325	4.0	0.02	1.3	4.0	0.0020	deV	3.0	0.85	dwarfs	BVI	Gilmore (1984, G84)
280	–	0.0028	1.9	–	–	–	–	–	dwarfs	BV	Tritton and Morton (1984, TM84)
200–475	–	0.016	1.18–2.21	–	0.0016	3.1	–	0.80	dwarfs, giants	UBV	Robin and Crézé (1986, RC86)
300	–	0.02	1.0	–	0.0010	deV	–	0.85	dwarfs	UGR	del Rio and Fenkart (1987, DF87)
285	–	0.015	1.3–1.5	–	0.0020	–	2.36	flat	dwarfs	UGR	Fenkart et al. (1987, F87)
325	–	0.0224	0.95	–	0.0010	deV	2.7	0.90	dwarfs, giants	UBV	Yoshii et al. (1987, Y87)
249	–	0.041	1.0	–	0.0020	–	3.0	0.85	dwarfs	BV	Kuijken and Gilmore (1989, KG89)
350	3.8	0.019	0.9	3.8	0.0011	deV	2.7	0.84	dwarfs, giants	UBV	Yamagata and Yoshii (1992, YY92)
290	–	–	0.86	–	–	deV	4.0	–	dwarfs, giants	uvby	von Hippel and Bothun (1993)
325	–	0.0225	1.5	–	0.0015	deV	3.5	0.80	dwarfs, giants	$B_J R_f$	Reid and Majewski (1993, RM93)

325	3.2	0.019	0.98	4.3	0.0024	–	3.3	0.48	2	dwarfs, giants	OE	Larsen (1996, L96)
250–270	2.5	0.056	0.76	2.8	0.0015	2.44	–	0.76	2	dwarfs, giants	UBV	Robin et al. (1996; 2000, R96, R00)
259	–	0.04	1.14	–	–	–	–	–	1	dwarfs	BVR _c	Spagna et al. (1996, Sp96)
250	–	0.053	1.0	–	–	3.0	–	1.0	2	dwarfs, giants	BV	Ng et al. (1997, N97)
290	4.0	0.059	0.91	3.0	0.0005	deV	2.69	0.84	1	dwarfs	UGR	Buser et al. (1998; 1999, B99)
240	2.5	0.061	0.79	2.8	–	–	–	0.60–0.85	1	dwarfs	UBV	Ojha et al. (1999, O99)
330	2.25	0.065–0.13	0.58–0.75	3.5	0.0013	2.5	–	0.55	2	dwarfs, giants	<i>u</i> <i>g</i> <i>r</i> <i>i</i> <i>z</i>	Chen et al. (2001, C01)
250	2.8	0.035	0.86	3.7	–	–	–	–	2	dwarfs, giants	JHK	Ojha (2001, O01)
280 (350)	2–2.5	0.06–0.10	0.7–1.0 (0.9–1.2)	3–4	0.0015	2.75	–	0.50–0.70	1	dwarfs	UBVRI	Siegel et al. (2002, S02, S02 _c)
320	–	0.07	0.64	–	0.0013	deV	–	0.58	1	dwarfs, giants	3000–10000Å	Du et al. (2003, D03)
–	3.5	0.006	0.9	4.7	0.0022	deV	4.3	0.55	2	dwarfs, giants	OE	Larsen and Humphreys (2003, LH03)
265–495	–	0.052–0.095	0.80–0.97	–	0.0002–0.0015	deV	–	0.70	1	dwarfs	<i>u</i> _{RGO} <i>g</i> _{RGO} <i>r</i> _{RGO} <i>i</i> _{RGO} <i>z</i> _{RGO}	Karaali et al. (2004, K04)
269	–	0.115	1.06	3.04	–	–	–	–	1	giants	JK	Cabrera-Lavers et al. (2005, C05)
300	2.6	0.04	0.9	3.6	0.005	2.8	–	0.64	1	dwarfs	<i>u</i> <i>g</i> <i>r</i> <i>i</i> <i>z</i>	Jurić et al. (2008, J08)
225	–	0.087	1.05	–	–	–	–	–	2	dwarfs, giants	JK	Veltz et al. (2008, V08)

Notes: *f* is the normalization relative to the thin disk near the sun, *h_z* and *h_g* are the exponential scale height and length, respectively, *r_e* is the effective radius, and *q* the axial ratio. The halo density law is either the de Vaucouleurs *r*^{1/4} law denoted “deV” or the *r*^{−*n*} law with the value of *n* given in the column. The method of analysis is either the photometric parallax analysis (PPA) denoted “1” or the model fit analysis (MFA) denoted “2.” The values in parentheses for Siegel et al. (2002) are the corrected values for binarity



■ Fig. 8-3

Previous star count results of the normalization f_{thick} plotted against the exponential scale height h_z for thick-disk stars. *Filled circles* and *squares* represent results based on model fit analysis (MFA) and photometric parallax analysis (PPA), respectively. The references besides the symbols are those in the last column of [Table 8-1](#). *Dotted lines* are given by $f_{\text{thick}} \times h_z^2 = \text{constant}$ for various constant values

measured since 1982 using star count studies, and [Fig. 8-3](#) shows plots of the exponential scale height h_z and normalization f_{thick} for the thick disk. After the last three decades of considerable improvement in both observations and modeling, some uncertainty regarding the structure of the thick disk still remains. Evidently, the derived values of f_{thick} and h_z are highly correlated with each other subject to the relationship $f_{\text{thick}} \times h_z^2 = \text{constant}$, the value of which gives the model's maximum contribution to the observed number of stars in a conic region surveyed in a certain direction. Various external constraints help distinguish f_{thick} from h_z and confine them in respective ranges of $f_{\text{thick}} \sim 0.01\text{--}0.1$ and $h_z \sim 0.5\text{--}2$ kpc.

Similar parameter coupling also appears for the halo, which is characterized by the power-law index n , axial ratio q , and normalization f_{halo} . Their derived values are highly correlated with each other subject to the relationship $f_{\text{halo}} \times q^2 \times n^{-n/2} (n-2)^{(n/2)-1} = \text{constant}$, for a reason similar to that in the case of the thick disk. It is interesting to plot the derived values of both q and f_{thick} in chronological order, as shown in [Fig. 8-4](#). Especially after 2000, there is a clear trend of larger reported normalizations of the thick disk, which coincides with a recent preference for a rather flattened halo having a smaller ratio of minor to major axes. Because the normalization

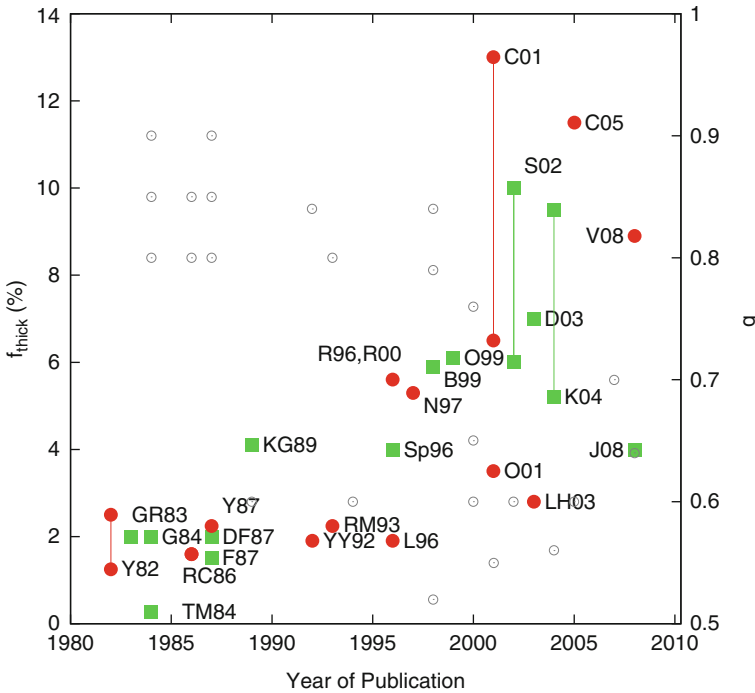


Fig. 8-4

Previous star count results of the thick-disk normalization f_{thick} (left ordinate) and the halo axial ratio q (right ordinate) plotted against the year of publication. Filled circles and squares represent results for thick-disk stars based on MFA and PPA, respectively. Open circles represent results for halo stars

of the halo is more or less fixed by dynamical constraints, a flattened halo necessarily predicts a smaller number of halo stars in high-latitude directions, with a compensating increase in the number of thick-disk stars. Accordingly, given that the structural parameters of the thin disk are well constrained, those of the thick disk seem to be more closely connected to the halo. To assess whether this connection is merely an artifact of star count analysis or in fact an evolutionary link between the thick disk and the halo, further evaluation of the large-scale spatial distribution of stars is necessary.

The chronological coincidence of reporting a larger thick-disk normalization and a smaller axial ratio of the halo corresponds to the availability of new deep high-quality data over a vast celestial coverage from recent high-latitude surveys, including the Sloan Digital Sky Survey (SDSS). Ongoing releases of these data having a wide sky coverage led to a renaissance of interest in detailed star count studies of the Galactic structure. One of the highlights of this stage was the successful determination of separate constraints on the exponential scale lengths of the thin and thick disks in the radial direction. These values are in the ranges of 2–2.5 kpc for a thin disk and 3–4 kpc for a thick disk, both of which are embedded in the more extended halo. It is extremely interesting to note that a larger-scale length for the thick disk is perfectly compatible

with recent results for other disk galaxies, which justifies the usual assumption of the Galaxy being one among many others (Yoachim and Dalcanton 2006).

The gradual decrease in the scale in order from the halo through the thick disk and eventually the thin disk might suggest that the observed properties of different stellar populations can be interpreted in terms of an evolutionary sequence beginning when galaxies formed. Along with efforts to refine the structural parameters of the thick disk, its origin was widely discussed (see an early review by Majewski 1993). For example, some considered the thick disk to be a completely separate population (Carney et al. 1989; Gilmore and Wyse 1985; Gilmore et al. 1989), while others saw it as a high-velocity tail of the thin disk or a low-velocity tail of the halo (Norris 1986; Norris and Green 1989). Progress in discriminating one from the other in this controversy required more insights into the distributions of the kinematics, chemical abundances, and ages of stars in the Galaxy. Thus, spectroscopic observations were thought to be essential for detailed studies of the dynamics and chemistry to supplement the multicolor photometric data, which could provide useful information on the gross metallicity and age as well as the spatial distribution of stars.

The following part of this chapter presents a method of star count analysis, reviews the present status and prospects of thick-disk studies based on a comprehensive survey of relevant data, and discusses several scenarios for its formation and evolution that have survived among numerous scenarios so far proposed. Thick-disk studies will be a key to understanding the fundamental features of the formation, evolution, and structure not only of the Galaxy but also of disk galaxies in general.

2 The Star Count Galaxy Model

It is no exaggeration to say that modern star count analysis began with the discovery of the double-peaked color distribution for high-latitude field stars at $20 < V < 22$, although the original motivation for the project was a cosmological interest in galaxy counts (Kron 1980; Tyson and Jarvis 1979). These studies identified the red and blue frequency peaks of the color distribution with the disk and halo components, respectively. Such stars at $V \sim 20$ mag in the blue peak can be at a distance of more than 10 kpc from the sun and reach the Galactic halo. Quite naturally, Bahcall and Soneira (1984) later showed that all the existing star count data in this magnitude range can be simulated by a Galaxy model with only two components, the disk and the halo.

It is remarkable that the double-peaked features persist down to $V \sim 16$ mag, where the Galaxy model shows too few halo stars to reproduce the blue peak, and hence thick disk stars must make up for this deficit. Therefore, the color distributions of stars at $V = 16\text{--}18$ mag are useful for constraining the structural parameters of the thick disk, in addition to a metallicity constraint primarily from the $U - B$ color data. Metal-poor stars, which are distributed farther away from the disk plane, generally have bluer colors; consequently, dwarfs have lower and giants have higher absolute luminosity compared to metal-rich stars of the same $B - V$ color. Therefore, the absolute magnitude, and hence the estimated distance, of stars at the same apparent magnitude depends strongly on the metallicity. This indicates that the space number density of distant stars must be solved in relation to the metallicity distribution in that region.

From the spectral analysis of bright stars, the observational calibration between the metallicity and the ultraviolet excess is known with high accuracy. The maximum color differences $\Delta(B - V) \sim 0.1$ mag and $\Delta(U - B) \sim 0.3$ mag at the turnoff point from the main sequence are

a result of the metallicity difference between $[Fe/H] = 0$ and -2 . In fact, the increase in the $\Delta(U - B)$ color difference with decreasing metallicity enables an accurate determination of the metallicity from the $U - B$ color distribution. In other words, to obtain a better understanding of the density and metallicity distributions of thick disk stars, accurate UBV photometry of a large number of field stars at $V = 16-18$ mag is indeed indispensable.

2.1 Fundamental Equation

The spatial distribution of stars in the Galaxy can be studied by counting stars along the line of sight away from the sun. The differential number of stars $A(m)$ counted as a function of apparent magnitude m over a solid angle ω is given by the fundamental equation of stellar statistics,


$$A(m) = \omega \int \phi(M)D(\mathbf{r})d^3\mathbf{r},$$

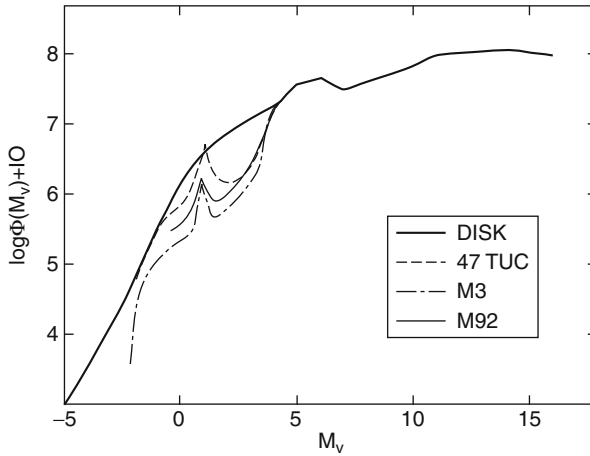
which is associated with the distance modulus equation,

$$m - M = 5 \log(r/10\text{pc}) + a(\mathbf{r}),$$

where \mathbf{r} is the position of the star along the line of sight from the sun, $\phi(M)$ is the LF for the number of stars with absolute magnitude M in cubic parsecs, $D(\mathbf{r})$ is the space number density function normalized to unity in the solar neighborhood, and $d^3\mathbf{r}$ is a volume element. The extinction $a(\mathbf{r})$ is 0.1 mag or less at high Galactic latitudes and is therefore unimportant for star-count analysis in such directions.

Actually, the fundamental equation is a linear sum of the contributions of $\phi_{ij}(M)D_i(\mathbf{r})$ over i and j in the integrand. The subscript i indicates the functions for the thin disk ($i = 1$), thick disk ($i = 2$), and halo ($i = 3$), whereas the subscript j refers to the main sequence ($j = 1$) and giant sequence ($j = 2$). Both of these sequences depend on $[Fe/H]$, which is a function of \mathbf{r} , where $[Fe/H] \equiv \log(Fe/H) - \log(Fe/H)_\odot$ is the metal abundance in solar units on a logarithmic scale. The i component is usually given a single density function regardless of whether the stars are on the main sequence or the giant sequence. If each (i, j) population is separated in a sample of star count data, the fundamental equation is powerful enough to determine $\phi(M)$ and $D(\mathbf{r})$ for each population with less uncertainty. However, because this pre-separation is generally not possible, some external constraints are applied in the star count model before the fundamental equation is solved. The result is therefore model dependent, and its reliability depends on that of the chosen inputs.

In this formulation, star count data can thus be simulated by integrating the LF $\phi(M)$ weighted by the space density of stars $D(\mathbf{r})$ with respect to a conic volume element $\omega d^3\mathbf{r}$ surveyed in a certain direction away from the sun. The fundamental equation contributes to an intuitive understanding of the relationship between $A(m)$ and the product of $\phi(M)D(\mathbf{r})$, but no unique segregation of this product by inversion is possible unless some other external constraints are imposed. The LF $\phi(M)$ is derived when the absolute magnitude M of the stars is evaluated reliably, either from trigonometric parallaxes of nearby stars or from a distance modulus common to member stars of a cluster. On the other hand, the density function $D(\mathbf{r})$ is adopted from the luminosity profiles of various external galaxies and is expressed by an analytic function, leaving a few numerical parameters to be determined. Such parameters for fixing $D(\mathbf{r})$ will be determined by reproducing $A(m)$ over the observed magnitude range with a realistic form of $\phi(M)$ (cf.  [Fig. 8-5](#)).



■ Fig. 8-5

Luminosity functions of nearby stars and of the globular clusters 47 Tuc, M3, and M92. The characteristic metallicities are, in order, $[Fe/H] = 0.0, -0.71, -1.67,$ and $-2.24,$ respectively. The globular cluster luminosity functions with the contribution of horizontal branch stars included are connected smoothly to that of nearby stars at $M_V = +4.$ The luminosity function, which varies as a function of galactocentric location, is obtained by interpolating between the four functions with respect to metallicity

The differential number of stars $A(m)$ is, without exception, known to monotonically increase as the apparent magnitude goes fainter in any high-latitude direction along the line of sight from the sun. The star counts at the brighter and fainter ends of the apparent magnitude are dominated by thin-disk and halo components, respectively, but no distinct feature of the thick disk appears in $A(m)$ at intermediate magnitudes in the transition from the brighter thin disk to the fainter halo.

On the other hand, it is remarkable that the frequency distribution of stars with respect to optical color $C,$ or the color distribution $A(m, C),$ shows a double-peaked feature. For example, in plotting stars in the V versus $B - V$ diagram, stars are densely populated at $B - V \sim 0.5$ and 1.4 at $V < 22$ mag. This is because the absolute magnitude M_V becomes very sensitive to $B - V$ at these specific color values; therefore, stars at various distances are observed in a very narrow $B - V$ range. This sensitivity of M_V to $B - V$ is clearly visible in the color-absolute magnitude (CM) diagram, on which the main and giant sequences are defined. The blue peak at $B - V \sim 0.5$ is populated by main-sequence stars plus subgiants near the turnoff that reach the thick disk or halo, depending on $m,$ well beyond the thin disk, whereas the red peak at $B - V \sim 1.4$ is populated mostly by late-type K and M main-sequence stars in the thin disk at distances of a few hundred parsecs from the sun. In particular, the red peak arises because the effective temperature and therefore the absolute magnitude are sensitive to optical color near $B - V \sim 1.4.$ In the CM diagram for the color of redder passbands, this sensitivity to the color becomes weaker; consequently, the red peak becomes lower for such cool stars.

Thus, by using the CM diagram, selecting the stars in a color bin is equivalent to assigning an absolute magnitude M and hence a distance modulus $(m - M)$ to each of these color-selected stars, which are on either the main sequence or the giant sequence. This explanation for the

emergence of the double-peaked feature justifies the use of the CM diagram as an external constraint on the fundamental equation.

When the CM diagram is used for stars in fields or globular clusters of appropriate metallicity, two methods are generally used to solve the fundamental equation: PPA and MFA. The former method assigns an absolute magnitude M to a subsample of main-sequence stars or giants binned at color C . Thus, in combination with the apparent magnitude m , this makes the determination of their distances straightforward and yields the product $\phi(M)D(r)$ for such color-selected stars of assumed luminosity class M . Extrapolation of $\phi(M)D(r)$ to $r = 0$ determines the density profile $D(r)$ for stars with M away from the solar neighborhood, as well as their normalization $\phi(M)$ at $r = 0$. Repetition over a range of C determines the solar-neighborhood LF $\phi(M)$. However, the feasibility of this method hinges upon how well main-sequence stars are separated from giants in the optical CM diagram or vice versa in the infrared CM diagram. Given that such color separation is not always complete, PPA is unavoidably subject to systematic bias.

The latter method models $\phi(M)$ upon external constraints in advance and determines the $D(r)$ that reproduces the differential number of stars $A(m)$ observed. MFA is obviously more model dependent, but full integration of the fundamental equation allows for the exploration of solutions in a wide parameter space. Note, however, that small changes in the modeled $\phi(M)$ cause corresponding changes of the parameters in $D(r)$. That is, the uncertainty in the normalization and scale size of the thick disk remains and may not be avoided, even if the prescriptions in the models are carefully checked for consistency with other available data on the kinematical and chemical properties of stars. Consequently, the difference in the reported results from author to author would be considered as some measure of unavoidable astrophysical uncertainty rather than a level of uncertainty in the analysis to be overcome in the future.

2.2 Input Data

The LF and color–absolute magnitude relation (CMR) for each subset of stars are known to differ depending on the metallicity, age, and other parameters, if any. Field stars in a unit space volume in the Galaxy obviously do not consist of a group of exactly coeval stars. Although age determination is not an easy task for individual field stars, their age is correlated to their metallicity as a result of the chemical evolution of the Galaxy. Thus, mainly for the purpose of simplicity, the metallicity is chosen as an explicit parameter because the average metallicity is correlated not only to the average age but also to the distance from the sun. Accordingly, like the CMR, the LF is also differentiated by $[\text{Fe}/\text{H}]$.

2.2.1 Solar Metallicity Inputs

The solar-neighborhood LF $\phi(M_V)$ for a wide range of absolute magnitudes has been derived from a volume-limited sample of nearby stars with reliable trigonometric parallaxes and proper motions (0.093 stars pc^{-3} for $M_V = -4$ to $+14$, McCuskey 1966; 0.11 stars pc^{-3} for $M_V = -1$ to $+20$, Wielen 1974; Wielen et al. 1983; 0.10 stars pc^{-3} for $M_V = -1$ to $+20$, Jahreiß and Wielen 1997). The results of Wielen and his colleagues (1974, 1983) were based

on Gliese's (1969) catalog of bright stars, and that of Jahreiß and Wielen (1997) was based on the Hipparcos catalog of nearby stars.

The LF shows a shallow frequency minimum at $M_V \sim +7$, rather than a smooth increase with M_V , as shown in [Fig. 8-5](#). The existence of this so-called Wielen dip, though less prominent in the improved determination by Jahreiß and Wielen (1997), has been confirmed (e.g., Reid et al. 2002). The turnover in $\phi(M_V)$ at $M_V \sim +12$ appears to be real, but the exact shape of the declining part at fainter magnitudes remains unsettled (e.g., Kroupa 1995; Reid and Gizis 1997; Reid et al. 2002). The fraction of nearby main-sequence stars at $M_V < +5$ is estimated from various sources compiled by Miller and Scalo (1979); their average is consistent with a later estimate by Jahreiß and Wielen (1997).

In the HR diagram, the color ($B - V$) to M_V relationship for nearby main-sequence stars is tight and has been defined to sufficient accuracy (Jahreiß and Wielen 1997; Perryman et al. 1995; Wielen 1974; Wielen et al. 1983), as shown in [Fig. 8-6](#). On the other hand, the distribution of nearby subgiants in the HR diagram is broad (Perryman et al. 1995), and the mean CMR is represented by the giant sequence of the Galactic open clusters NGC 188 or M67, with the lower envelope of their distribution set by the giant sequence of NGC 6791 (Sandage et al. 2003). Thus, the usual practice for using the mean CMR of subgiants through red giants is to adopt the

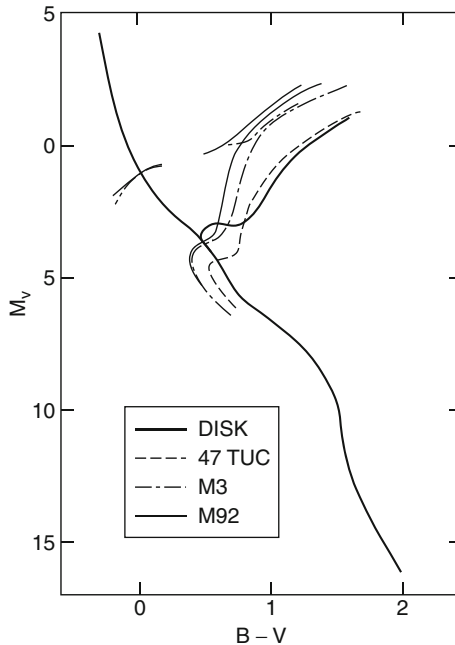


Fig. 8-6

Color ($B - V$) to M_V relationship for nearby main-sequence and giant stars, and those of the globular clusters 47 Tuc, M3, and M92. The relationship for nearby giant stars is taken from the giant sequence of the old Galactic cluster M67 and is connected to that of nearby main-sequence stars at the turnoff magnitude. The color–magnitude relationship, which varies as a function of galactocentric location, is obtained by interpolating between the four relationships with respect to metallicity

giant sequence of M67 in the magnitude range brighter than its turnoff magnitude $M_V(\text{TO})$. The basic data given in Sandage et al. (2003) for these clusters are ($[\text{Fe}/\text{H}]$, $M_V(\text{TO})$, age) = (0.00, 4.15, 6.2 Gyr) for NGC 188, (-0.04, 3.85, 4.0 Gyr) for M67, and (+0.37, 4.50, 10.0 Gyr) for NGC 6791.

2.2.2 Lower Metallicity Inputs

Our knowledge of the LF and CMR for metal-poor field stars is still incomplete, and the current practice in star count studies is to use those of the metal-poor globular clusters as templates having different metallicity values, which in turn depend on the distance from the sun through the spatial metallicity gradient (Bahcall and Soneira 1980; Yoshii 1982).

This practice has been supported by some observational trends. By using distant stars in the SGP region and assuming the known metallicity gradient in this direction, the LF $\phi(M_V)$ at $M_V < +4$ was found to have a steeper slope with increasing vertical distance up to a few kiloparsecs above the Galactic plane, similar to the aging effect seen in globular clusters (Gilmore and Reid 1983). Furthermore, by using high-velocity stars in the solar neighborhood, the LF of metal-poor field stars for the halo component was first determined by Schmidt (1975) and later refined by Bahcall and Casertano (1986) and Dawson (1986). Their results show that the slope of $\phi(M_V)$ at $M_V < +4$ is much steeper than Wielen's function, and the overall shape of $\phi(M_V)$ in the magnitude range of $+5 < M_V < +12$ agrees more or less with Wielen's function scaled by a factor of 1/800–1/600, although the exact shape of $\phi(M_V)$ at fainter magnitudes is still under debate (Dahn et al. 1995; Digby et al. 2003; Gould 2003). The LF $\phi(M_V)$ of metal-poor stars would then be constructed by smoothly connecting the brighter part ($M_V < +4$) of the globular cluster's function to the fainter part ($M_V > +4$) of Wielen's function at $M_V = +4$.

Regarding the CMR, metal-poor main-sequence stars are bluer and brighter because of line blanketing and eventually form a subdwarf sequence below the main sequence of metal-rich stars. By deriving the accurate distances and absolute magnitudes of nearby stars in the Hipparcos program, metal-poor stars were found to form separate subdwarf sequences of different metallicities that agree with globular cluster isochrones of the corresponding metallicities (Bartkevicius et al. 1997; Gizis 1997; Reid 1997).

The globular clusters typically adopted as templates of different metallicities are 47 Tuc ($[\text{Fe}/\text{H}] = -0.71$, LF: Da Costa 1982; CMR: Hesser et al. 1987), M3 ($[\text{Fe}/\text{H}] = -1.67$, LF: Sandage 1957; Simoda and Fukuoka 1976; CMR: Rood et al. 1999), and M92 ($[\text{Fe}/\text{H}] = -2.24$, LF: Fukuoka and Simoda 1976; Hartwick 1970; Paust et al. 2007; CMR: Stetson and Harris 1988). Although the metallicity scale for these clusters varies slightly from author to author, it is customary to use the compilation of Zinn and West (1984). For 47 Tuc, a lower value of $[\text{Fe}/\text{H}] = -0.85$ was reported by Gratton and Ortolani (1989), but their later analysis based on high-resolution spectroscopic observations returned the value to the original Zinn and West scale (Carretta and Gratton 1997; Gratton et al. 2003). Other basic data for these clusters are ($M_V(\text{TO})$, age) = (+4.2, 11 Gyr) (Gratton et al. 1997, 2003) for 47 Tuc, (+4.0, 12.5 Gyr) (Rood et al. 1999; VandenBerg et al. 2000) for M3, and (+3.9, 14 Gyr) (Gratton et al. 1997; VandenBerg et al. 2000) for M92.

The data for these three clusters and those of solar-neighborhood stars form a four-point grid of metallicity in the range of $-2.24 \leq [\text{Fe}/\text{H}] \leq 0.0$. Thus, the LF and CMR for any metallicity in this range are obtained by interpolation of the data with respect to $[\text{Fe}/\text{H}]$. This can of course be extended by increasing the number of template clusters and/or by adding more

parameters of interest (e.g., age, kinematics, elemental abundance ratio) as extra dimensions. However, the four-point grid is fairly standard in star count studies and suffices for practical purposes.

2.3 Functional Forms

A Galaxy model is then constructed by solving the remaining numerical parameters in the density and metallicity distributions. The purpose is to examine whether the Galaxy has an appreciable amount of intermediate thick-disk component in addition to the thin disk and halo and how steep the metallicity gradient is in the direction perpendicular to the disk plane.

In practice, because both the density and metallicity distributions are observed from the sun, it is important to fix the solar position in the Galaxy model. The typically adopted radial distance R_\odot of the sun from the Galactic center is 8 kpc (Reid 1993), but the solar offset or its vertical displacement Z_\odot north of the midplane of the disk has been estimated by many authors to date. These estimates, based on the north–south difference in the solar-neighborhood normalization, lie in a range from 15 pc (Wolf–Rayet stars, Conti and Vacca 1990; IRAS source counts, Cohen 1995; surface brightness map, Binney et al. 1997) to 40 pc (high-latitude star counts, Yamagata and Yoshii 1992) and currently seem to converge on $Z_\odot = 25$ pc (Jurić et al. 2008 and references therein).

2.3.1 Density Function

Assuming the structure of the Galaxy exhibits rotational symmetry and north–south symmetry overall, the density law for stars usually adopted for both thin and thick disks is a double exponential function in cylindrical RZ coordinates,

$$D(\mathbf{r}) = f \exp[-(R - R_\odot)/h_R - |Z + Z_\odot|/h_Z],$$

where f is the normalization in the solar neighborhood, and h_R and h_Z are the scale sizes assigned to old, late-type dwarfs in the radial and vertical directions, respectively. Whereas the thin disk is normalized to unity by definition, the thick disk is normalized to the solar-neighborhood thin disk. This form is purely empirical.

The thick-disk component was first identified in some S0 galaxies (Burstein 1979) and other disk galaxies (van der Kruit 1984; van der Kruit and Searle 1981, 1982). Deprojection of their surface brightness profiles is the only way to infer the functional form of the large-scale thick-disk density law reliably from the outside, and it is reasonable to consider the thick disk as having the same density function as the thin disk but with different normalization and scale sizes.

The density law is also expressed by using a similar but theoretically well-defined form of a vertical disk,

$$D(\mathbf{r}) = f \exp[-(R - R_\odot)/h_R] \operatorname{sech}^2[-|Z + Z_\odot|/(2h_Z)],$$

where $h_Z \equiv \sigma_w^2/(2\pi G\Sigma)$, G is the gravitational constant, Σ is the mass density in the vertical column, and σ_w is the vertical W -velocity dispersion. This vertical distribution is an exact solution of the Poisson equation for a self-gravitating isothermal sheet, and no discontinuity in

the force emerges across the midplane of the disk (van der Kruit and Searle 1981). The sech^2 function leads to the Gaussian for $|Z| \ll h_Z$ and the exponential for $|Z| \gg h_Z$. These limiting cases are suggestive when a more realistic distribution is to be explored.

Which of the two forms of $D(r)$ above is preferred for the fits to optical star counts still seems inconclusive (Gould et al. 1996; Pritchett 1983), but the exponential has been claimed to yield superior fits to near-infrared star counts (Hammersley et al. 1999). Moreover, the near-infrared surface brightness distribution of edge-on galaxies shows, without hindrance by dust obscuration, that the luminosity Z profile is best fitted by the exponential all the way to near the midplane of the disk (de Grijs et al. 1997). Given a recent trend in favor of the pure exponential in the vertical direction, other functional forms including the sech^2 function are examined only to estimate the range of uncertainty in the disk parameters to be determined.

The density law of the thick disk must be treated with caution. Because the difference between the exponential and sech^2 functions is small and the predicted number of thick-disk stars is easily hidden in the larger number of dominant thin-disk stars near the disk plane; star count analysis alone cannot distinguish the exponential from the alternative. Instead, an appropriate choice of the functional form requires kinematical information for stars that are presumably in the thick disk. If their Z velocity dispersion remains constant with no vertical gradient, the isothermal sech^2 function is preferred. On the other hand, if their Z velocity dispersion increases monotonically away from the disk plane, the exponential distribution with a single scale height may be an appropriate choice.

The density law for a spheroidal halo is expressed well by the deprojected form of the de Vaucouleurs $r^{1/4}$ profile of the surface brightness distribution of elliptical galaxies (Young 1976):

$$D(r) = f(s/R_\odot)^{-7/8} \exp[-br_e^{-1/4}(s^{1/4} - R_\odot^{1/4})].$$

The argument s is defined as

$$s^2 = R^2 + (Z + Z_\odot)^2/q^2 \sim R^2 + Z^2/q^2,$$

where q is the axial ratio of the equidensity surface, $b = 7.669$ is a numerical constant, and r_e is the distance projected to an observed angle within which half of the total brightness is enclosed. In the case of the Galaxy, r_e is nearly equal to $R_\odot/3$ (de Vaucouleurs and Pence 1978). The deprojected form successfully yields a color distribution that peaks at $B - V \sim 0.5$, in agreement with high-latitude star count data.

A variety of power-law density functions are often used for the spheroidal halo, such as $1/s^n$, $1/(a+s)^n$, $1/(a^n + s^n)$, $1/[s(a+s)^{n-1}]$, and so on, where a is the core radius. For a power index $n = 3-4$, these forms give a density profile very similar to de Vaucouleurs $r^{1/4}$ profile, especially at high Galactic latitudes. Because all of these forms are empirical and there is no particular reason to select one over the others, de Vaucouleurs $r^{1/4}$ profile suffices to account for high-latitude star count data.

The derived values of q and n based on the overall distribution of various halo tracers include the results of $q = 0.7-1$ and $n = 3-4$ (metal-poor globular clusters, Bica et al. 2006; Harris 1976; RR Lyrae variables, Hawkins 1984; Vivas and Zinn 2006; blue horizontal branch stars, Brown et al. 2008; Sommer-Larsen and Christensen 1989). However, because n and q are coupled to each other in star count analysis, the possibility of an extended, flatted density distribution with $n < 3$ and $q < 0.7$ must be explored for field halo stars.

2.3.2 Metallicity Function

A simple metallicity distribution having three linear segments in the cylindrical RZ coordinates is adopted:

$$[\text{Fe}/\text{H}] = \begin{cases} 0 + \alpha_Z|Z + Z_\odot| + \alpha_R(R - R_\odot) & \text{for } |Z + Z_\odot| < Z_{\text{thick}}, \\ \beta + \beta_Z|Z + Z_\odot| & \text{for } |Z + Z_\odot| \geq Z_{\text{thick}}, \\ \gamma & \text{for } |Z + Z_\odot| \geq Z_{\text{halo}} \end{cases}$$

where α , β , and γ represent the thin disk, thick disk, and halo, respectively, and the estimated boundaries of these components are $Z_{\text{thick}} \sim 1\text{--}2$ kpc and $Z_{\text{halo}} \sim 5\text{--}7$ kpc (e.g., Ivesić et al. 2008). The subscript R represents the radial gradient $d[\text{Fe}/\text{H}]/dR$, and the subscript Z represents the vertical gradient $d[\text{Fe}/\text{H}]/d|Z|$.

Note that the LF and CMR of the thick disk and halo are not given in advance but are derived as a result of assigning appropriate metallicities to their respective components through MFA. The α component yields a mean thin-disk metallicity of $[\text{Fe}/\text{H}] \sim -0.3$ with a rather large vertical gradient up to Z_{thick} (e.g., Du et al. 2004). The β component yields a thick-disk metallicity of $[\text{Fe}/\text{H}] \sim -0.8$ or less at around Z_{thick} and becomes dominant up to Z_{halo} with essentially no metallicity gradient $\beta_R \approx 0$ (e.g., Allende Prieto et al. 2006). The γ component yields a halo metallicity of $[\text{Fe}/\text{H}] \sim -1.5$ (e.g., Ivesić et al. 2008).

Given $|\alpha_Z| > |\beta_Z| \sim |\alpha_R| \sim |\beta_R| \approx 0$ (e.g., Allende Prieto et al. 2006; Ivesić et al. 2008), distance determination for high-latitude stars is most sensitive to the vertical metallicity distribution of the α component, particularly its gradient α_Z . Estimates by various spectroscopic observations fall within a range of -0.2 dex kpc^{-1} to -0.8 dex kpc^{-1} and accumulate at ~ -0.5 dex kpc^{-1} (GK giants, Hartkopf and Yoss 1982; Yoss et al. 1987; FG dwarfs, Gilmore and Wyse 1985; Trefzger et al. 1983, 1995). A simple mean of $\alpha_Z = -0.5$ dex kpc^{-1} is a reasonable prescription in the Galaxy model.

2.3.3 Model Parameters

In summary, the number of parameters for the density distribution is eight: the thin-disk scale sizes (h_R , h_Z), thick-disk normalization and scale sizes (f_{thick} , h_R , h_Z), and halo normalization, axial ratio, and power-law index (f_{halo} , q , n). The number of parameters for the metallicity distribution is five: the thin-disk metallicity gradients (α_R , α_Z), thick-disk metallicity intercept and vertical gradient (β , β_Z), and halo metallicity γ . These parameters can be determined by applying the least-squares method to various high-latitude star count data. In practice, however, the number of such parameters, chosen so as to be determined by external constraints on the others, varies from author to author, which makes a fair comparison among the numerous thick-disk parameters in the literature difficult.

In comparing observations with the predictions of star count models, several limitations apply. Star count data in a few directions close to the Galactic poles, although best studied, scarcely constrain the radial density distribution of any component. On the other hand, star count data in various directions away from the poles can in principle constrain the radial density distribution, but unless the directions are carefully chosen, local enhancement of stellar streams would fictitiously flatten the density distribution, notably for the halo. The assumption

of rotational and north–south symmetries for the structure of the Galaxy may be too idealistic, which could partially explain the failure to reproduce the halo star count data in various directions simultaneously by the model.


In most cases, the parameters for both the thin and thick disks are fitted to star count data with the halo parameters prescribed in advance. In fact, the thick-disk parameters to be derived are coupled with those of the thin disk and inherently with those of the halo as well. Furthermore, in the near-polar directions there exist well-known couplings of $f_{\text{thick}} \propto h_Z^{-2}$ for the thick disk and $f_{\text{halo}} q^2 \propto n^{n/2} (n-2)^{-(n/2)+1}$ for the halo, where n is the index of the power-law density function of the halo. Thus, MFA of the fundamental equation inevitably produces some range of descriptions of the thick-disk structure, and a near-unique description must break such component coupling by star count observations in various off-polar directions and by spectroscopic follow-up observations of the kinematics, age, and chemical abundance of potential in situ candidates of thick-disk stars at a distance of $|Z| \sim 2h_Z$.

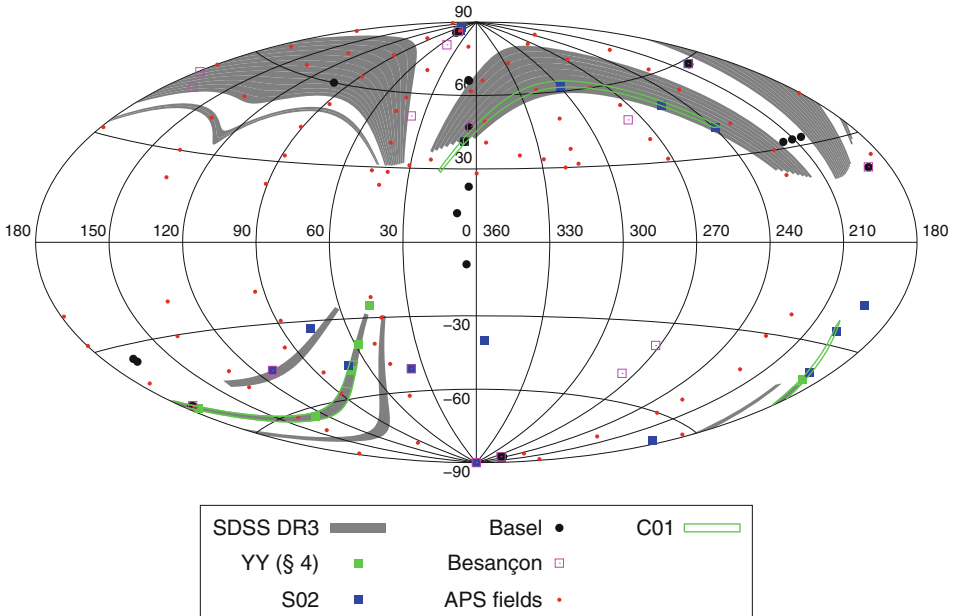
2.4 Star Count Observations

As described above, either PPA or MFA, applied to the right-hand side of the fundamental equation, determines the density function and the metallicity/age-dependent LF. The reliability of the derived functions, however, depends on what type of star count data will be used on the left-hand side of the fundamental equation. In fact, counts of field stars in several passbands to a sufficiently faint limiting magnitude over large celestial areas would be most suitable for providing essential information on the structure of the Galaxy.

Modern star count studies have proceeded as several technological milestones were reached. The first milestone, which was reached in the 1980s, was triggered by the invention of the automated plate measuring machine. This enabled a large number of field stars to be counted over a full area $\sim 20 \text{ deg}^2$ on Schmidt plates to a typical limiting magnitude of $V_{\text{lim}} \sim 18\text{--}19 \text{ mag}$ (SGP in 18.24 deg^2 in VI [UK Schmidt], Gilmore and Reid 1983; $(l, b) = (37^\circ, -51^\circ)$ in 17 deg^2 in BVI [UK Schmidt], Gilmore et al. 1985; NGP in 21.46 deg^2 in UBVI [Kiso Schmidt], Yoshii et al. 1987; SA 54 ($200^\circ, 59^\circ$) in 16 deg^2 in UBVI [Kiso Schmidt], Yamagata and Yoshii 1992; NGP in 19.9 deg^2 , M5 ($2.7^\circ, 47.4^\circ$) in 15.5 deg^2 , anticenter ($167.5^\circ, 47.4^\circ$) in 7.13 deg^2 , M3 ($50^\circ, 80^\circ$) in 20.26 deg^2 , antirotation ($277.8^\circ, 46.7^\circ$) in 20.84 deg^2 in UBVI [Tautenburg Schmidt and other telescopes], Robin et al. 1996). These wide-survey data greatly improved the statistics such that density fluctuations of stars in each direction became less important.

At the same time, CCDs realized higher photometric precision and fainter limiting magnitudes but with an obvious drawback of a much smaller area in each direction just after their initial availability. Thus, especially in star count studies, real acquisition of CCD data was long delayed (NGP in 0.158 deg^2 and $(l, b) = (150^\circ, 60^\circ)$ in 0.051 deg^2 in VI at $V < 20 \text{ mag}$, Robin et al. 2000; seven SA fields [141, 101, 102, 107, 184, 90, 114] in a total of 14.9 deg^2 in BVRI at $V < 21.4 \text{ mag}$, Siegel et al. 2002; the ($169.95^\circ, 49.80^\circ$) area in 0.95 deg^2 in 15 intermediate bands at $V < 21$, Du et al. 2003).

The next milestone, which occurred recently, was triggered by the exclusive use of the 2.5-m SDSS survey telescope. In fact, as shown in  Fig. 8-7, continual release of SDSS multicolor *ugriz* star count data in the 2000s achieved unprecedented sky coverage far exceeding that of any other CCD survey conducted so far (north [151 deg^2] and south [128 deg^2] samples for $49^\circ < |b| < 64^\circ$ and $g < 21 \text{ mag}$, Chen et al. 2001; north [$5,450 \text{ deg}^2$] and south [$1,088 \text{ deg}^2$] samples for



■ Fig. 8-7

Target fields of major star count programs in the Aitoff projection. Shown are the fields used by Basel and Besançon programs, Siegel's CCD fields (S02), and the selected fields of APS catalog used by Larsen and Humphreys (2003). The SDSS early data in two stripes 10 and 82 were used by Chen et al. (2001, C01) and the SDSS third data release (DR3) data over the large area were used by Jurić et al. (2008). SDSS stripe 82 data in several selected fields (green squares) were used to calculate the color distributions in ● Figs. 8-11 and ● 8-12

a wider range of $|b|$ and $g < 22$ mag, Jurić et al. 2008). Because they were processed using the same calibration and reduction procedures, these data exhibit excellent homogeneity.

Regardless of whether the surveys are deep in small areas or shallow over wide areas, obtaining star counts in only a few high-latitude directions inevitably results in degenerate determination of the density law. This degeneracy is in principle removed by analyzing the data from multidirection surveys covering a range of l and b over the sky. Following this motivation, roughly a dozen fields were strategically selected over the sky in several major survey programs, such as the Basel program by Becker and his colleagues [UGR with $G < 19$ mag in 14 fields of $2\text{--}3\text{ deg}^2$ each] (Becker 1980; Buser et al. 1998), the Besançon program by Robin and her colleagues [UBV with $V < 18$ mag in five fields of $10\text{--}20\text{ deg}^2$ each, plus other authors' fields] (Robin et al. 1996), and Siegel's CCD program [$BVRI$ with $V < 21.4$ mag in seven SA fields of $2\text{--}3\text{ deg}^2$ each, plus other SA fields with incomplete RI data] (Siegel et al. 2002).

In addition, star count data covering more than 10^3 deg^2 are now available from far more extensive survey catalogs, such as the APS catalog, which provides all-sky optical data in the O (blue) and E (red) passbands with $O < 20$ mag by scanning the original Palomar glass plates for 88 selected fields of 16 deg^2 each (Larsen and Humphreys 2003); the 2MASS catalog, which provides all-sky near-infrared JHK data with $K < 11$ mag in all the 2MASS fields in 15° around the Galactic poles (Cabrera-Lavers et al. 2005); and

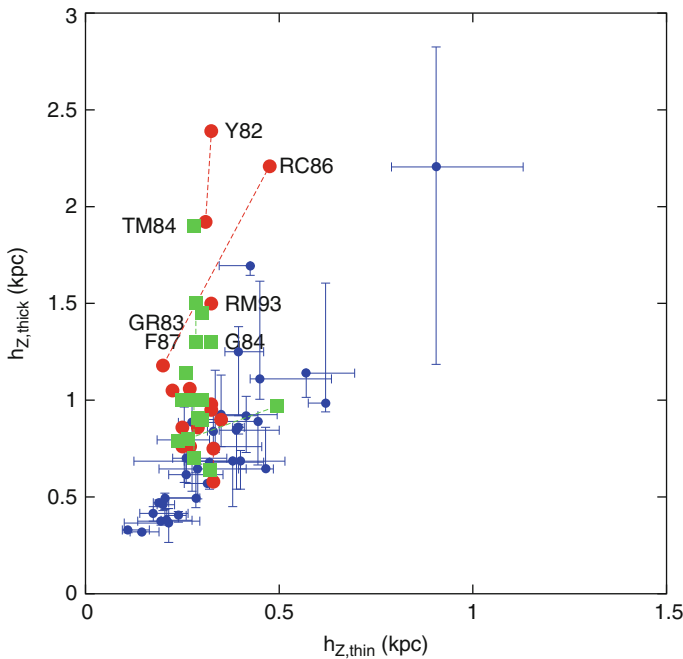
the SDSS third data release (DR3) catalog, which provides optical *ugriz* data with $g < 22$ mag in long, wide stripes on the sky (Jurić et al. 2008).

3 Structural Constraints from Star Counts

3.1 Estimates of Structural Parameters

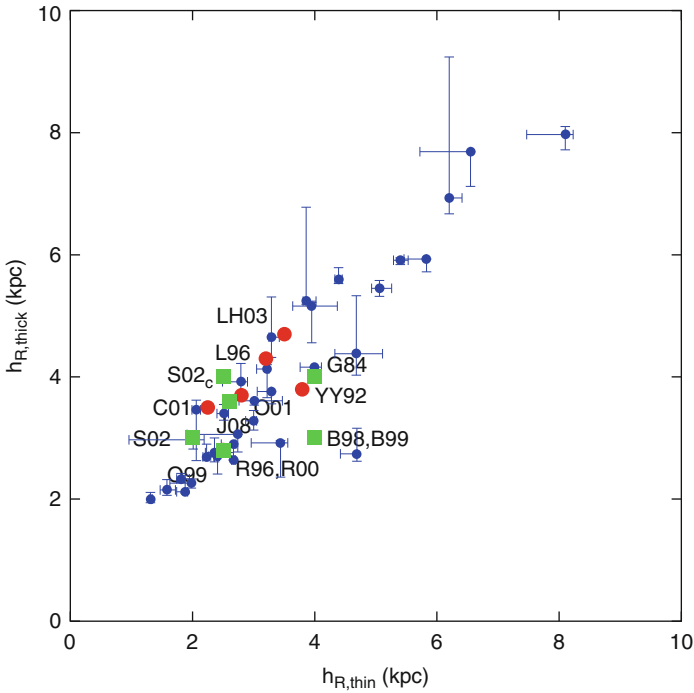
► *Table 8-1* tabulates the star count results in the literature in chronological order. The disk structure parameters that result from halo prescriptions of $f_{\text{halo}} = 1/800\text{--}1/600$ and $q > 0.5$ lie in a range of $h_Z = 250\text{--}350$ pc and $h_R = 2\text{--}4$ kpc for the thin disk, and $f_{\text{thick}} = 2\text{--}10\%$, $h_Z = 0.7\text{--}1.5$ kpc, and $h_R = 2.5\text{--}4$ kpc for the thick disk (► *Figs. 8-8* and ► *8-9*).

There are three noteworthy points here. First, the recently reported thin-disk structure has smaller values of $h_Z \sim 250\text{--}280$ pc and $h_R \sim 2\text{--}2.5$ kpc, compared with previous estimates of $h_Z \sim 300\text{--}325$ pc and $h_R \sim 3.5\text{--}4$ kpc. Second, the thick-disk structure follows $f_{\text{thick}} \propto h_Z^{-2}$ and exhibits a recent trend toward a larger normalization of $f_{\text{thick}} \sim 0.07\text{--}0.1$ or a smaller height of $h_Z \sim 0.7\text{--}1$ kpc, compared with previous estimates of $f_{\text{thick}} \sim 0.02\text{--}0.04$ or $h_Z \sim 1\text{--}1.5$ kpc, while retaining $h_R = 3.5\text{--}4$ kpc. Third, the recently reported halo structure has a more extended



■ Fig. 8-8

Previous results of exponential thick-disk scale height $h_{Z,\text{thick}}$ plotted against thin-disk scale height $h_{Z,\text{thin}}$. Filled circles and squares show the results of Galactic star counts based on MFA and PPA, respectively. Small dots represent those from the luminosity profile decomposition of external disk galaxies by Yoachim and Dalcanton (2006), by converting their Z_0 into $h_Z (\equiv Z_0/2)$



■ Fig. 8-9

Previous results of exponential thick-disk scale length $h_{R,thick}$ plotted against thin-disk scale length $h_{R,thin}$. Others are same as ● Fig. 8-8

and more flattened distribution with $n \sim 2-2.5$ and $q \sim 0.5-0.6$, compared with the previous estimates of $n \sim 3-4$ and $q > 0.8$.

These changes in the Galactic structure described in the literature occurred, not over a long period, but somewhat suddenly in 1997–2002 after the Galaxy was surveyed in multiple directions to great depth far into the halo. This sudden transition is attributed to a preference for an extended, flattened halo that necessitates a preference for the vertical concentration of the thick disk as well as the thin disk.

It is possible that the recent transformation of the star count results is an artifact owing to parameter coupling during least-squares fitting procedures. Therefore, the result should be checked to see whether it is consistent with external results other than star count analysis.

3.2 Implications of Recent Estimates

A smaller-scale height of $h_z \sim 250-280$ pc for the thin disk has been reported in many recent star count studies (● Table 8-1). The previous understanding here, according to Miller and Scalo (1979), is that the scale height increases steeply from $M_V \sim +3$ to $M_V \sim +5$, beyond which a constant scale height (300–325 pc) is given by extrapolation. However, this switchover at $M_V \sim +5$ is uncertain, and the consequent smaller h_z suggests that the increase in the scale height is more

gradual at $M_V \sim +5$ and continues to a fainter magnitude than that usually adopted. Because the scale height has a one-to-one correspondence with the vertical velocity dispersion, the thin disk may not be a single component but would be made up of several subcomponents of old stars having different velocity dispersions.

A smaller-scale length of $h_R \sim 2\text{--}2.5$ kpc for the thin disk has also been reported in near-infrared star count studies (Ojha 2001; Ortiz and Lepine 1993; Porcel et al. 1998; Ruphy et al. 1996) and agrees well with COBE DIRBE observations of the spatial distribution of Galactic dust, which is a tracer of ongoing star formation in the disk (Drimmel and Spergel 2001). The scale ratio for edge-on disk galaxies was estimated as $h_R/Z_0 = 5.9 \pm 0.4$ (de Grijs and van der Kruit 1996), but this was later refined to $h_R/Z_0 = 4 \pm 0.5$ for Sc-type galaxies similar to the Galaxy (de Grijs 1998). The application of this refined ratio to the previous standard value of $h_R \sim 3.5\text{--}4$ kpc yields $h_Z \sim 440\text{--}500$ pc ($h_Z \equiv Z_0/2$), which is far beyond any higher-scale heights of the Galactic thin disk in the literature. With $h_R \sim 2\text{--}2.5$ kpc, however, the resulting $h_Z \sim 250\text{--}310$ pc is reconciled with recent estimates of h_Z for the Galactic thin disk.

A smaller axial ratio of $q \sim 0.5\text{--}0.6$ for the spheroidal halo is associated with a preference for an extended density law with a power index of $n \sim 2\text{--}2.5$ (Table 8-1). The reason is that because the halo normalization f_{halo} has been estimated by dynamical constraints (Bahcall and Casertano 1986; Dawson 1986; Schmidt 1975), little room remains to use it as a free parameter. Therefore, results of $n \sim 2\text{--}2.5$ and $q \sim 0.5\text{--}0.6$ are obtained with the halo normalization almost fixed but not left to be determined. This oblate halo from recent star count studies agrees with the axial ratio of the spatial distribution of metal-rich globular clusters but is not easily reconciled with their steep density decline with $n \sim 3.9$ (Bica et al. 2006). The star count result for the oblate halo, which was first claimed by Wyse and Gilmore (1989), is also consistent with a recent claim of an inner oblate halo embedded in the surrounding near-spherical outer halo (Carollo et al. 2007, 2010), although this dual-halo paradigm is in dispute (Schönrich et al. 2011).


3.3 Vertical Scale Height of Thick Disk

Extrapolation of the thick-disk density derived from star counts at $|Z| \sim 1\text{--}2$ kpc yields thick-disk normalization in a range of $f_{\text{thick}} = 0.01\text{--}0.1$ of the local density. Although most estimates cluster around $f_{\text{thick}} = 0.02\text{--}0.06$, some recently reported estimates are as high as $f_{\text{thick}} \sim 0.1$. Star count analysis alone may not be able to determine which is more plausible, and a further constraint on f_{thick} in the solar neighborhood would be obtained from dynamical arguments by measuring the W -velocity dispersion of candidate thick-disk stars at $|Z| \sim 2\text{--}4$ kpc.

The results for f_{thick} in SDSS star count studies are worthy of remark. Chen et al. (2001) applied the MFA method to data from stripes 10 and 82 and derived $f_{\text{thick}} = 0.065\text{--}0.13$, which is higher than the more recent estimate of $f_{\text{thick}} = 0.04$ of Jurić et al. (2008) based on the PPA method and the DR03 catalog, assuming the main-sequence stars as a tracer population. The difference in their results cannot be attributed solely to their use of different methods. As shown in Fig. 8-3, various estimates of f_{thick} and h_Z by the PPA method span the same region in the $f_{\text{thick}}\text{--}h_Z$ diagram as that covered by MFA estimates, and no systematic difference appears between the regions covered by each method. If more credence is given to the results of far more extensive multidirection surveys, the lower normalization of Jurić et al. (2008) would be preferable, although it is not compatible with the post 2000 trend in favor of a higher normalization of $f_{\text{thick}} \sim 0.1$. In particular, extinction-free infrared star count data from the all-sky 2MASS survey yield a higher normalization of $f_{\text{thick}} = 0.115$, by applying the PPA method to red clump stars

separated on the K versus $J - K$ diagram (Cabrera-Lavers et al. 2005). The most recent result, based on a much smaller sample taken from the same 2MASS catalog, gives a similar result of $f_{\text{thick}} = 0.087$ (Veltz et al. 2008).


Comparing thick-disk normalizations reported by various authors is difficult because their PPA studies do not always use the same tracer population and those using MFA do not use exactly the same assumptions and prescriptions in the Galaxy model. Furthermore, as Siegel et al. (2002) pointed out, many authors normalize the thick disk relative to the old thin disk without any contributions from young stars, which necessarily overestimates the true normalization to be compared with dynamical constraints.

Many estimates of the thick-disk scale height, except for obvious outliers, lie in a range of $h_Z \sim 0.7\text{--}1.2$ kpc. This range is somewhat narrow, unlike the reported range of f_{thick} , because h_Z is less sensitive to changes in f_{thick} , reflecting the near-vertical behavior of $f_{\text{thick}} \propto h_Z^{-2}$, as shown in  Fig. 8-3. Estimates of h_Z from near-polar and nonpolar surveys do not differ. Thus, with the current density laws of the thin disk and halo, it is certain that the thick disk with $h_Z \sim 0.7\text{--}1.2$ kpc dominates at $|Z| \sim 1\text{--}2$ kpc in stellar number and further extends up to $|Z| \sim 5\text{--}7$ kpc (e.g., Reid and Majewski 1993).

3.4 Radial Scale Length of Thick Disk

Multidirection surveys are the only way to examine the radial density law of the thick disk. The lower the latitude surveyed, the more reliably the radial scale length is determined by comparing the star counts in the directions toward the Galactic center with those in the anticenter directions. In earlier studies, the thick-disk scale length was not constrained to sufficient accuracy and was at best assumed to be equal or comparable to that of the thin disk. It is only recently that extensive sets of high-quality star count data have enabled separate determinations to yield somewhat convergent results for a thick-disk scale length of $h_R = 3$ kpc (Basel data, Buser et al. 1998, 1999), $h_R = 2.8$ kpc (Besançon data, Robin et al. 1996, 2000), $h_R = 3\text{--}4$ kpc (CCD multidirection data, Siegel et al. 2002), $h_R = 2.8 \pm 0.3$ kpc (2MASS data, Ojha 2001), and $h_R = 3.04 \pm 0.11$ kpc (2MASS data, Cabrera-Lavers et al. 2005).

In one exception, a much larger thick-disk scale length of $h_R = 4.7 \pm 0.2$ kpc was reported by Larsen and Humphreys (2003) on the basis of all-sky optical star count data. This may have arisen partly from inadequate correction for field-to-field variations in extinction. In fact, extinction-free analysis yields $h_R = 2.8\text{--}3.0$ kpc based on all-sky near-infrared 2MASS star count data (Cabrera-Lavers et al. 2005; Ojha 2001) and $h_R = 3.5 \pm 0.5$ kpc based on all-sky 2MASS-selected blue horizontal branch stars (Brown et al. 2008), in agreement with other estimates of $h_R \sim 3$ kpc.

More importantly, however, Larsen and Humphreys' scale length of the thick disk is significantly larger than their own estimate of the scale length of the thin disk, which reinforces a similar conclusion regarding the Galaxy (Cabrera-Lavers et al. 2005; Jurić et al. 2008; Ojha 2001; Robin et al. 1996; Siegel et al. 2002). This trend of a larger-scale length for thick-disk stars than for thin-disk stars is clearly visible in  Fig. 8-9.

Determination of the global structure of the Galaxy is a difficult task in itself. In particular, the scale length of the disk, among other parameters, is somewhat weakly constrained by multidirection star count analysis within a distance not very far from the sun. As an independent check, it would be worth asking where the current estimates of Galactic structure fit in comparison with those of other disk galaxies.

Yoachim and Dalcanton (2006) decomposed the optical R -band surface brightness distribution into two components corresponding to thin and thick disks for a sample of 34 disk galaxies viewed edge-on. First, they confirmed the universal existence of a thick disk, which indicates that the thick disk in the Galaxy is not an accidental emergence but an inevitable consequence of the processes of galaxy formation. Second, they found that the scale height h_Z ($\equiv Z_0/2$) and length h_R for the thin and thick disks increase almost linearly with the circular velocity V_c , or equivalently the mass of the galaxy, and that this trend is fully consistent with the results derived from the surface brightness distribution of nearby edge-on S0 galaxies (Pohlen et al. 2004) and the resolved stellar populations in nearby edge-on disk galaxies (Seth et al. 2005). Their values of h_Z and h_R at $V_c \sim 220 \text{ km s}^{-1}$ are systematically larger than those of the Galaxy, probably owing to dust obscuration of the optical light, analogous to the larger scales obtained in optical star counts in the Galaxy by Larsen and Humphreys (2003).

On the other hand, the scale ratios likely escape this systematics and give more reliable estimates. \bullet [Figures 8-8](#) and \bullet [8-9](#) clearly show $h_{Z,\text{thick}}/h_{Z,\text{thin}} \sim 2$ and $h_{R,\text{thick}}/h_{R,\text{thin}} \sim 1.2$ with some scatter of about 20–30%, in good agreement with those of the Galaxy, the Andromeda Galaxy (Collins et al. 2011), and nearby edge-on S0 galaxies (Pohlen et al. 2004). Thus, the larger scales of the thick disk in both the radial and vertical directions, compared with the thin disk, should be taken as real, which suggests more strongly than ever that the thin and thick disks may have emerged from a self-regulating disk formation mechanism.

3.5 Systematics in the Results

3.5.1 PPA or MFA

According to the usual practice in which all stars in a sample are assumed to be on the main sequence, the PPA method is liable to give a density distribution with a systematically smaller-scale size for both thin and thick disks, even if the correction for metallicity is properly applied. This decrease in scale size is necessarily associated with an increase in thick-disk normalization to keep the predicted number of stars unchanged along the line of sight (Gilmore and Reid 1983; Yoshii 1982; Yoshii et al. 1987).

The reliability of PPA results depends on how reasonable the assumption of no giant contamination is. Stars spend a significant fraction of time just after the turnoff as subgiants and on the ascending branch as red giants. These two types of giants are worthy of consideration. Because the ascending branch of red giants is nearly vertical in the CM diagram, the effect of red giant contamination is minimized by excluding stars in the color range of the ascending branch, but this exclusion has the obvious disadvantage of simultaneously excluding late dwarfs of the same color. Except for this specific color range, however, an absence of red giant contamination becomes a good assumption for star counts at faint magnitudes ($V > 17 \text{ mag}$) and high latitudes ($|b| > 50^\circ$), where the metal-poor stars of the thick disk dominate. This is because red giants are much brighter and distant and therefore dominate only at sufficiently bright apparent magnitudes and/or toward sufficiently low-latitude directions.

On the other hand, neither the color nor the apparent magnitude can distinguish subgiants near the turnoff. Thus, subgiant contamination is unavoidable, and the results of PPA must be corrected. Estimates of this correction require the product of $\phi(M)D(r)$ for both subgiants

and main-sequence stars near the turnoff, which is essentially equivalent to the MFA procedure. [Table 8-1](#) shows that the range of scale sizes estimated by PPA is similar to that covered by MFA. Given this similarity, the systematic difference, if any, may not be very significant in the end.

3.5.2 Binary Effect

A large fraction of nearby disk stars are known to be binaries, with estimates varying from 57% (G dwarfs, Duquennoy and Mayor 1991), 35–42% (M dwarfs, Fischer and Marcy 1992; Reid and Gizis 1997), down to 33% (F–K dwarfs, Raghavan et al. 2010). The apparent angular size of binaries with a separation $a \sim 100$ AU, if placed at a distance $r \sim 100$ pc from the sun, is $\sim 1'' (a/100 \text{ AU})/(r/100 \text{ pc})$, which is comparable to a typical seeing in current imaging surveys. Thus, most binaries beyond the solar neighborhood are unresolved and are counted as single stars. In other words, the apparent magnitudes of such single stars are brightened; therefore, their distances are underestimated. Estimates of this bias are difficult and depend on not only the binary fraction but also the relative luminosity L_1/L_2 and effective temperature T_1/T_2 between primary and secondary stars.

Pedagogical simulations based on simple distributions of L_1/L_2 and T_1/T_2 by Siegel et al. (2002) showed that a binary fraction of 50% decreases the scale height of the density distribution by about 20% for both the thin and thick disks. However, a recent preference for a smaller fraction ($\sim 30\%$) for nearby dwarfs (e.g., Raghavan et al. 2010; Reid and Gizis 1997) and an even smaller fraction (10–20%) for metal-poor subdwarfs (e.g., Riaz et al. 2008) suggests that the effect of binaries on the derived density scales would be smaller than so far discussed.

4 Interpretation of the Hess Diagram

For illustrative purposes, the Galaxy model of Yoshii and his colleagues (Yamagata and Yoshii 1992; Yoshii et al. 1987) was run using recently updated parameter values, as listed in [Table 8-2](#). The sample of field stars to be compared with the model was taken from the SDSS *ugriz* catalog for equatorial stripe 82 in the south Galactic plane, specifically, six different fields of 10 deg^2 each separated by about 10° in latitude along stripe 82. A wide l and b coverage, with accurate photometry (~ 0.02 mag) and faint flux ($r < 22$ mag), enables us to examine how much the current settings of parameter values succeed in modeling the large-scale stellar distribution of the Galaxy.

The distribution of stars in the g versus $g - r$ diagram or the Hess diagram for a combined sample of stars in these six fields is shown in [Fig. 8-10](#), which is the same as Fig. 6a in Chen et al. (2001), except that not all the stars in the entire stripe are plotted here and the correction for small extinction (0.02–0.03 mag) is not applied. Multiple ridges are apparent, and their existence is a function of g . A brighter ridge of blue color at $g - r \approx 0.4$ is prominent down to $g \sim 19$. Two fainter ridges at $g - r \sim 0.3$ and 1.4, which are already visible at $g \sim 18$, coexist and persist to $g > 21$. According to the model, the brighter ridge at $g - r \sim 0.4$ is populated by the thin disk at $g < 17$ and then dominated by the thick disk to $g \sim 19$. The fainter ridges at $g - r \sim 0.3$ and 1.4 are populated exclusively by the halo and thin disk, respectively. Therefore, the brighter blue ridge of the thin disk at $g - r \sim 0.4$, after passing through $g \sim 17$, connects to the fainter red ridge of

■ Table 8-2

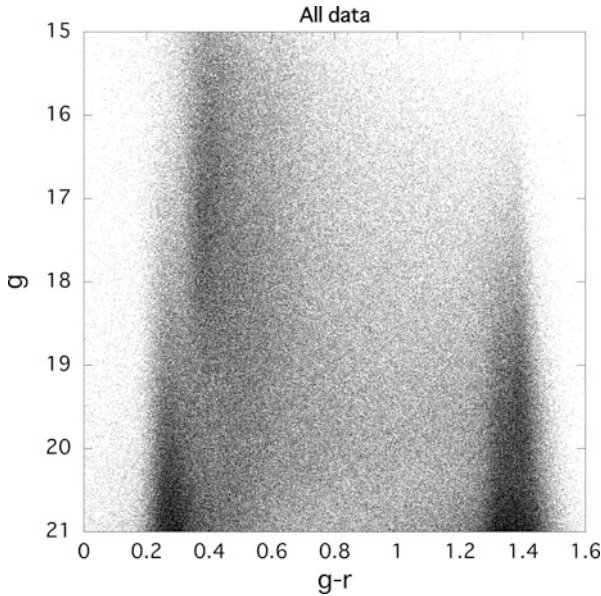
Parameters of the star count Galaxy model

Solar position		
Galactocentric distance		8 kpc
Distance from the disk plane		25 pc north
Thin disk component		
Local normalization		0.093 stars pc ⁻³
Scale height	Late-type dwarfs	300 pc
	Subgiants	250 pc
Scale length		2.5 kpc
Luminosity function		McCluskey (dip)
CM diagram	Main sequence	Wielen
	Giant sequence	Open cluster M67
Metallicity at $(R, Z) = (R_{\odot}, 0)$		0.0 dex
Metallicity gradient,	Z direction	-0.5 dex kpc ⁻¹
	R direction	-0.04 dex kpc ⁻¹
Thick disk component		
Local normalization to disk		0.1/0.02
Scale height		0.7/1 kpc
Scale length		3.5 kpc
Luminosity function		Globular clusters
CM diagram		Globular clusters
Mean metallicity		-0.8 dex
Metallicity gradient,	Z direction	-0.1 dex kpc ⁻¹
Halo component		
Local normalization to disk		0.00125 (=1/800)
Density law		r^{-n} power law
Power index		2.5/3.5
Axial ratio		0.55/0.8
Luminosity function		Globular clusters
CM diagram		Globular clusters
Mean metallicity		-1.5 dex

Current combinations of parameters consist of an extended thick disk (f_{thick}, h_z) = (0.02, 1 kpc) plus a near-spherical halo (n, q) = (3.5, 0.8) as in [Fig. 8-11](#), and a compact thick disk (0.1, 0.7 kpc) plus an oblate halo (2.5, 0.55) as in [Fig. 8-12](#)

the thin disk at $g - r \sim 1.4$, whereas the blue ridge of the thick disk at $g - r \sim 0.4$ barely appears to $g \sim 20$ and dissolves into the growing fainter blue ridge of the halo at $g - r \sim 0.3$. For reference, the colors of $g - r \sim 0.3, 0.4,$ and 1.4 correspond to $B - V \sim 0.4, 0.5,$ and 1.5 , respectively.

The change in the dominant blue ridge in the intermediate range of $18 < g < 19$ is associated with the change in the turnoff color from the thick disk to the halo. Thus, the color distribution, particularly around the blue ridge, places internally consistent constraints on the metallicities of the thick disk and halo as well as their ages on the basis of metallicity-dependent templates



■ Fig. 8-10

The g versus $g - r$ diagram for SDSS sample of stars in equatorial stripe 82 in the south Galactic plane. The same as in Chen et al. (2001), but for different fields of 10 deg^2 each separated by about 10° in latitude along stripe 82

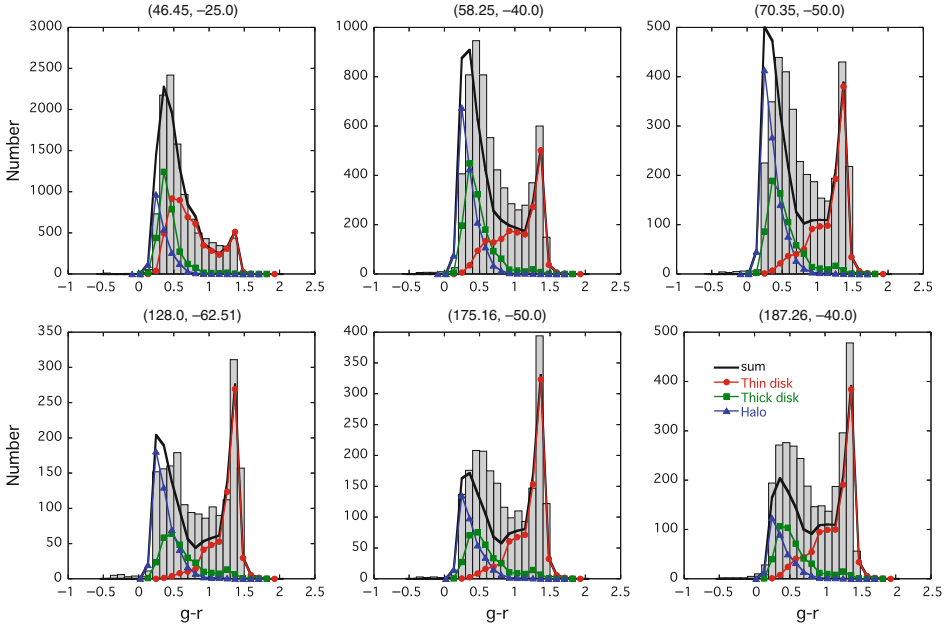
of globular clusters (Chen et al. 2001; Gilmore and Reid 1983; Reid and Majewski 1993; Siegel et al. 2009; Yoshii et al. 1987).

The double-peaked $g - r$ color distributions of stars in six selected fields are shown in [Fig. 8-11a](#) for $18 < g < 19$ and in [Fig. 8-11b](#) for $20 < g < 21$. An extended thick disk of $(f_{\text{thick}}, h_z) = (0.02, 1 \text{ kpc})$ is used here, with a near-spherical halo of $(n, q) = (3.5, 0.8)$. Predicted counts in color bins of 0.1 mag are represented by thick lines as a sum of the contributions of the thin disk (circles), thick disk (squares), and halo (triangles). Observed counts are presented as a histogram. Because the size of the color bin is set larger than the uncertainties of the photometry, the step-like blue edge at $g - r \sim 0.2$ is real.

For $20 < g < 21$, the predicted and observed color distributions in the different fields considered agree roughly. However, for $18 < g < 19$, the predicted color position of the blue ridge is slightly bluer than that observed in the fields of $|b| \geq 50^\circ$. Thus, the problem here is only with the thick disk, not the halo. Among other possible parameter settings, a model of a compact thick disk embedded in an oblate halo is intriguing; consequently, the color distributions with $(f_{\text{thick}}, h_z) = (0.1, 0.7 \text{ kpc})$ and $(n, q) = (2.5, 0.55)$ were calculated. The results are shown in [Fig. 8-12a](#) for $18 < g < 19$ and in [Fig. 8-12b](#) for $20 < g < 21$. The agreement is obviously better, although it is still not perfect, for $18 < g < 19$; in contrast, for $20 < g < 21$, this oblate halo is prone to underpredict the blue counts at $|b| < 30^\circ$.

Thus, each model has its own problems at different magnitudes and cannot account for the color distributions over the full magnitude range down to $g \sim 21$ mag. This problem could be solved by using a hybrid model consisting of a near-spherical outer halo and an oblate inner halo that surrounds a compact thick disk. This dual halo is consistent with recent chemical and kinematical studies based on a subsample of SDSS stars by Carollo et al. (2007, 2010).

a $18 < g < 19$



b $20 < g < 21$

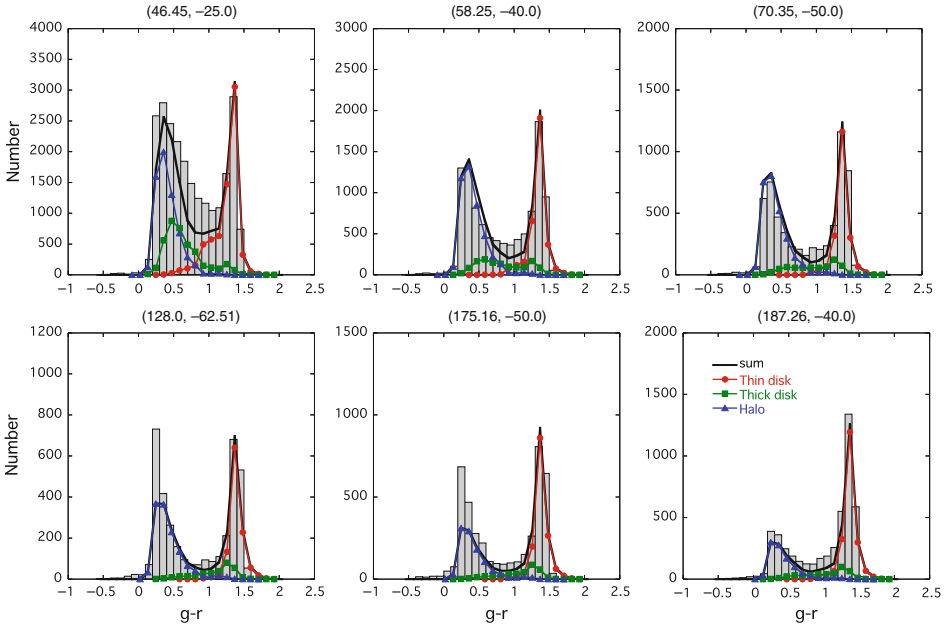
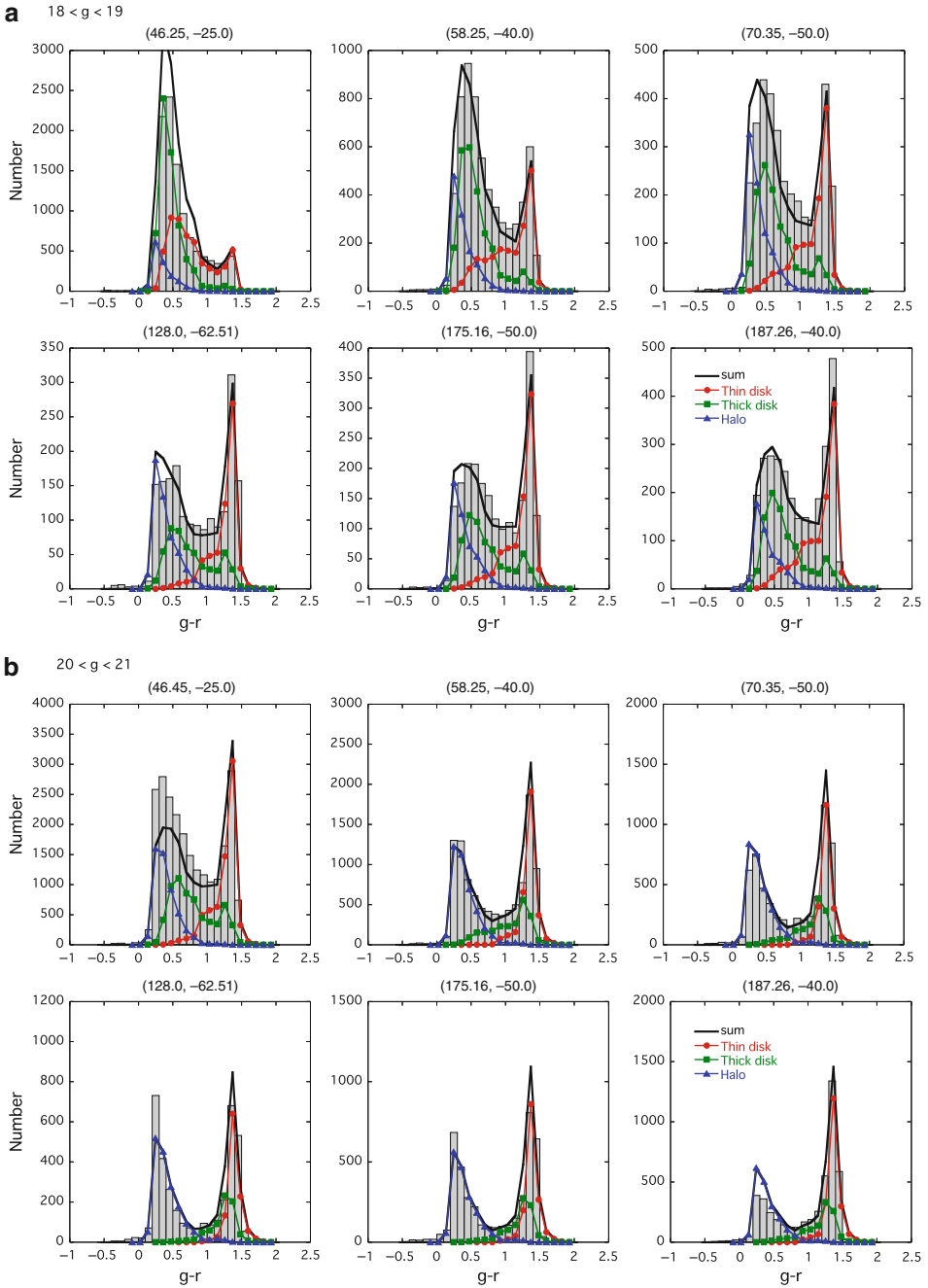


Fig. 8-11

Model color distributions for stars in six selected fields (l, b) for an extended thick disk of $(f_{\text{thick}}, h_z) = (0.02, 1 \text{ kpc})$ and a near-spherical halo of $(n, q) = (3.5, 0.8)$. Upper and lower six panels are for (a) $18 < g < 19$ and (b) $20 < g < 21$, respectively. Contributions from thin disk, thick disk, and halo are indicated by *thin lines with circles (red), squares (green), and triangles (blue), respectively*; their sum is indicated by the *thick line (black)*. Histogram shows star count data



■ Fig. 8-12

Model color distributions for stars in six selected fields (l, b) for (a) $18 < g < 19$ and (b) $20 < g < 21$. Same as Fig. 8-11 but for a compact thick disk of $(f_{\text{thick}}, h_z) = (0.1, 0.7)$ kpc and an oblate halo of $(n, q) = (2.5, 0.55)$

A closer look reveals that the extended thick-disk model for $18 < g < 19$ (► *Fig. 8-11a*) underpredicts the blue counts in directions toward the anticenter of the Galaxy while providing reasonable fits in the quadrants toward the Galactic center. Alternatively, the compact thick-disk model for $18 < g < 19$ (► *Fig. 8-12a*) overpredicts the blue counts at $|b| < 30^\circ$ toward the Galactic center while providing reasonable fits toward the anticenter of the Galaxy. Changing the thick-disk normalization fixes the problem toward the anticenter or the center but obviously causes a problem in the opposite direction. This seesaw-like behavior, which may have arisen from the usual assumption of a constant scale height, suggests flaring of the thick disk in which the scale height is either larger toward the anticenter (► *Fig. 8-11a*) or smaller toward the center (► *Fig. 8-12a*). However, it remains to be confirmed whether the thick-disk density distribution should be modified as above (e.g., Lopez-Corredoira 2006) or a thick disk having a constant scale height with a suitably adjusted normalization would still be a solution.

Metal-poor stars are intrinsically bluer and brighter than metal-rich stars of the same mass. Thus, the metallicity gradient along the line of sight from the sun systematically effects the resulting distances of stars, from which their density distribution is derived. Because the color position of the blue ridge corresponds to the turnoff color, which depends on the metallicity and age, the fits to the observed color distribution constrain the density and metallicity distributions simultaneously in a self-consistent manner.

The metallicity distributions for different components can then be deduced from the features of the blue ridges in the g versus $g - r$ diagram. Their clear verticality at $g - r \sim 0.4$ for the thick disk and $g - r \sim 0.3$ for the halo indicates that the metallicity gradients, if any, for these components are very small with typical mean metallicities of $[\text{Fe}/\text{H}] \sim -0.8$ for the thick disk and -1.5 for the halo. In addition, the verticality at $g - r \sim 0.4$ evidently holds from brighter magnitudes ($g < 17$), where the metal-rich stars of the thin disk dominate, to fainter magnitudes ($g > 17$), where the metal-poor stars of the thick disk dominate. No noticeable color change at $g - r \sim 0.4$ appears during this transition of the dominant components, which requires a nonzero metallicity gradient for the thin disk that should guarantee a smooth transition of the thin-disk metallicity to that of the thick disk.

The metallicity-gradient range could be estimated in terms of the color uncertainty δC within which the apparent verticality of the blue ridge holds, for example, $\Delta[\text{Fe}/\text{H}]/\Delta C(\text{TO}) \times |\delta C| \div \Delta d$, where $\Delta[\text{Fe}/\text{H}]$ is the metallicity change associated with a change in the turnoff color by $\Delta C(\text{TO})$, and Δd is the range of distance over which a considered component is dominant. Substitution of $\Delta[\text{Fe}/\text{H}]/\Delta C(\text{TO}) \sim 4\text{--}5 \text{ dex mag}^{-1}$ from 47 Tuc and M3, $\Delta d \sim 3\text{--}4 \text{ kpc}$ in the vertical direction, and $|\delta C| < 0.05 \text{ mag}$ for $g > 17$ from visual inspection yields $|\beta_Z| < 0.1 \text{ dex kpc}^{-1}$ for the thick disk.

In practice, as shown in ► *Figs. 8-11a* and ► *8-12a*, reasonable fits to the observed blue counts for $18 < g < 19$ would justify the current setting of $\alpha_Z = -0.5 \text{ dex kpc}^{-1}$ for the thin disk and $\beta_Z = -0.1 \text{ dex kpc}^{-1}$ for the thick disk. These metallicity gradients were previously derived using MFA of the photographic color distribution in $B - V$ as well as the metallicity-sensitive $U - B$ in the NGP region (Yoshii et al. 1987). However, a thick disk with no vertical metallicity gradient yields a least-squares sum which is only slightly worse than that with a gradient of $-0.1 \text{ dex kpc}^{-1}$ (Yamagata and Yoshii 1994), reflecting the fact that the color distribution used does not have sufficient power to distinguish such a small but finite vertical metallicity gradient from the complete absence of a gradient.

Siegel et al. (2009) revisited this problem using their more precise CCD-based UBV star count data in the SGP region of SA 141. The ultraviolet excess is used to estimate the metallicity, and the photometric parallax is used to derive the distance of individual stars, assumed to be on

the main sequence for their estimated metallicity. The vertical gradient of the mean metallicity detected at $|Z| < 4$ kpc is approximately -0.15 dex kpc^{-1} and is best explained by a star count model that allows a spatially varying numerical ratio of metal-rich thin-disk stars to metal-poor thick-disk stars that have a mean of $[\text{Fe}/\text{H}] = -0.7$ with a vertical gradient as small as -0.05 dex kpc^{-1} .

Recent extensive analyses of the spectroscopic/photometric metallicities and photometric parallaxes of Galactic field stars in the massive SDSS database have provided accurate metallicity distributions for a huge number of stars at distances out to ~ 30 kpc from the sun. Allende Prieto et al. (2006) constructed a spectroscopic metallicity distribution for several thousands of candidate thick-disk stars of FG spectral types at $1 \text{ kpc} < |Z| < 3 \text{ kpc}$ included in SDSS DR3. They derived a median value of $[\text{Fe}/\text{H}] = -0.68$ and a 1σ dispersion of 0.24 dex with an apparent lack of metallicity gradients in both the vertical and radial directions. They remarked that the vertical gradient for the thick disk, if present, must be shallower than -0.03 dex kpc^{-1} .

Ivesić et al. (2008) analyzed the proper motions and photometric metallicities of several tens of thousands of FG stars in SDSS DR6 and found a non-Gaussian disk metallicity distribution, with a trend in the median metallicity to decrease from $[\text{Fe}/\text{H}] = -0.6$ dex at $|Z| = 0.5$ kpc to -0.8 dex beyond 3 kpc. The vertical metallicity gradient of -0.1 dex kpc^{-1} detected in this range of distance from the disk plane is more or less consistent with previous results (Allende Prieto et al. 2006; Siegel et al. 2009) but is interpreted differently as the *real* gradient within a single disk component instead of the *apparent* gradient produced by the transition from a metal-rich thin disk to a metal-poor thick disk.

The interpretation of Ivesić et al. (2008) implies the absence of a distinct thick-disk component, whereas their data clearly confirm the slope change in the number density law at $|Z| \sim 1$ kpc by which the thick disk was first demonstrated. This is because their data do *not* show a significant correlation between the metallicity and the rotational velocity distributions for stars at $|Z| > 1$ kpc, which must be confirmed if the thick disk is really a distinct component. In contradiction to Ivesić et al. (2008), however, Spagna et al. (2010) recently found evidence of a metallicity–rotation correlation based on more accurate proper motions derived from a new data release, SDSS DR7.

Nevertheless, it has become evident that the verticality of the blue ridge at $g - r \sim 0.4$ in the g versus $g - r$ diagram requires a vertical metallicity gradient of -0.5 dex kpc^{-1} away from the disk plane that gradually flattens to maintain $[\text{Fe}/\text{H}] = -0.7 \sim -0.8$ dex over to $|Z| \sim 4$ kpc. Whether the thick disk has a zero or nonzero metallicity gradient is still unclear, but its vertical gradient seems to be shallower than -0.1 dex kpc^{-1} .

In summary, the mean metallicity in space is characterized by a smoothly decreasing function with respect to distance from the Galactic disk plane. Furthermore, the observed metallicity gradient near the plane is not a result of the decreasing numerical ratio of metal-rich disk stars to metal-poor stars of the other component with increasing distance from the plane.

5 Other Constraints on the Thick Disk

As described, the scale height of the thick disk is coupled with other structural parameters, and contrary to early expectations, it cannot be determined uniquely even by multidirection surveys. Thus, this coupling remains as a methodological weakness in star count analysis, and its removal requires external constraints.

5.1 Kinematical Constraints

For any disk-like system in dynamical equilibrium, the vertical structure is related to its vertical dynamics to the degree that plane-parallel stratification is a reasonable approximation of the mass distribution in the Z direction. Thus, the vertical motion of stars is analogous to an oscillation across the midplane of the disk for which the disk gravity determines the vertical W -velocity dispersion σ_w .

For example, a tracer population that follows an exponential density law with scale height h_Z is expected to have σ_w values at $|Z| \sim h_Z$ resembling $\sigma_w^2 \approx 2\pi G \Sigma h_Z$, where Σ is the surface mass density of the disk (van der Kruit 1988). A sophisticated model of star counts combined with the vertical dynamics yields $\sigma_w \sim 40\text{--}50 \text{ km s}^{-1}$ to be consistent with the thick-disk scale height of $h_Z \sim 0.7\text{--}1 \text{ kpc}$, whereas it yields a larger σ_w for a larger h_Z (e.g., Just and Jahreiß 2010; Ojha 2001; Veltz et al. 2008). Thus, direct measurement of σ_w for the thick disk could break the coupling and constrain the range of h_Z .

Stars located at $|Z| \sim 1\text{--}3 \text{ kpc}$ are presumably potential candidates for a thick disk between the thin disk and the halo. In fact, many studies have directly measured σ_w for such stars, and their results range from 30 to 50 km s^{-1} (35 km s^{-1} , Norris 1986; 42 km s^{-1} , Sandage 1987; $47 \pm 5 \text{ km s}^{-1}$, Carney et al. 1989; $30\text{--}35 \text{ km s}^{-1}$, Croswell et al. 1991; $38 \pm 4 \text{ km s}^{-1}$, Beers and Sommer-Larsen 1995; $35 \pm 3 \text{ km s}^{-1}$, Chiba and Beers 2000). Apparently, a range of $\sigma_w \sim 30\text{--}50 \text{ km s}^{-1}$ corresponds to $h_Z \sim 0.5\text{--}1 \text{ kpc}$, giving dynamical credence to a shorter scale height of the thick disk below 1 kpc .

In addition to σ_w in the Z direction, the other two velocity dispersions in axisymmetrical coordinates are the radial U -velocity dispersion σ_u and the rotational V -velocity dispersion σ_v . Some results for $(\sigma_u, \sigma_v, \sigma_w)$ for kinematically unbiased samples of stars give a useful measure of the velocity dispersion anisotropy of the thick disk ($75, 35, 42 \text{ km s}^{-1}$, Sandage and Fouts 1987; $63 \pm 7, 42 \pm 4, 38 \pm 4 \text{ km s}^{-1}$, Beers and Sommer-Larsen 1995; $46 \pm 4, 50 \pm 4, 35 \pm 3 \text{ km s}^{-1}$, Chiba and Beers 2000; $65 \pm 2, 39 \pm 9, 41 \pm 2 \text{ km s}^{-1}$, Alcobé and Cubarsi 2005; $70 \pm 4, 48 \pm 8, 36 \pm 4 \text{ km s}^{-1}$, Casetti-Dinescu et al. 2011). These results converge somewhat, except for Chiba and Beers (2000), and a comparison between them and the result for the thin disk (e.g., $28 \pm 1, 16 \pm 2, 13 \pm 1 \text{ km s}^{-1}$, Alcobé and Cubarsi 2005) indicates that the two disks share a common feature of a velocity ellipsoid elongated in the radial direction ($\sigma_u/\sigma_v \sim \sigma_u/\sigma_w \sim 1.5\text{--}2$), despite the difference in their velocity values. Interestingly, the ratio of stellar kinematic energy per unit mass (other than rotational energy) between the two disks does not depend greatly on the direction. In other words, this energy change, either from the thick disk to the thin disk or vice versa, should have occurred without any especially preferred direction.

Information on systematic motion is contained in the mean U , V , and W -velocities, and the condition of rotational equilibrium ($\langle U \rangle \ll \sigma_u, \langle W \rangle \ll \sigma_w, \langle V \rangle \gg \sigma_v$) is satisfied for the thick disk. Although $\langle W \rangle$ is marginally consistent with 0 km s^{-1} , as expected from north-south symmetry, $\langle U \rangle$ is reported to deviate from the expected value of 0 km s^{-1} by a few kilometers per second beyond plausible systematic errors. A small, positive value of $\langle U \rangle$, if maintained over the age of the thick disk, suggests its secular outward evolution, as discussed in terms of a possible radial migration of stars in the disk. The mean rotational velocity $\langle V \rangle$ of the azimuthal streaming motion of the thick disk is estimated as $\sim 180\text{--}170 \text{ km s}^{-1}$. This lags by $\sim 40\text{--}50 \text{ km s}^{-1}$ behind the rotational velocity at a local standard of rest $V_{\text{LSR}} = 220 \text{ km s}^{-1}$ (e.g., Beers and Sommer-Larsen 1995; Spagna et al. 2010) and is sometimes expressed relative to V_{LSR} , that is, $\langle V_{\text{lag}} \rangle \equiv \langle V \rangle - V_{\text{LSR}}$. (Note, however, that the so-called asymmetric drift is defined as $V_{\text{LSR}} - \langle V \rangle$.)

The three components of the velocity dispersion ($\sigma_u, \sigma_v, \sigma_w$) and the mean rotational velocity $\langle V \rangle$ representing the thick disk are intermediate between those of the thin disk and the halo. Thus, a question arises as to whether the thick disk is dynamically discrete from the other two or is connected to them by a continuous trend. A quantification of the vertical gradient in the kinematics from the disk plane may help answer this question.

A vertical gradient in $\langle V \rangle$ for the thick disk was detected a few decades ago (see the review by Majewski 1993), but its value is still uncertain, depending on the survey region and size as well as the correction for thin-disk contamination in the sample. Some estimates prefer a linear Z gradient of $\sim -20 \text{ km s}^{-1} \text{ kpc}^{-1}$ or shallower ($-21 \pm 1 \text{ km s}^{-1} \text{ kpc}^{-1}$, Majewski 1992; $-14 \pm 5 \text{ km s}^{-1} \text{ kpc}^{-1}$, reanalysis of Majewski's data by Chen 1997; $-16 \pm 4 \text{ km s}^{-1} \text{ kpc}^{-1}$, Allende Prieto et al. 2006; $-19 \pm 2 \text{ km s}^{-1} \text{ kpc}^{-1}$, Spagna et al. 2010), while other estimates prefer a much steeper gradient ($-30 \pm 3 \text{ km s}^{-1} \text{ kpc}^{-1}$, Chiba and Beers 2000; $-30 \pm 3 \text{ km s}^{-1} \text{ kpc}^{-1}$, Girard et al. 2006; $-25 \pm 2 \text{ km s}^{-1} \text{ kpc}^{-1}$, Casetti-Dinescu et al. 2011). Whichever is the case, the vertical rotation gradient should be considered one of the important constraints for discriminating between possible thick-disk formation scenarios.

The vertical gradients of the three components of the velocity dispersion have been reliably derived only recently. This is because large amounts of kinematical data have become available, and the statistical errors have decreased significantly. For example, Girard et al. (2006) analyzed a thousand red giants in the SGP region and derived gradients of $7.5 \pm 3.1 \text{ km s}^{-1} \text{ kpc}^{-1}$ in σ_u and $10.5 \pm 3.3 \text{ km s}^{-1} \text{ kpc}^{-1}$ in σ_v from a subset of stars at $|Z| = 1\text{--}4 \text{ kpc}$. On the other hand, Casetti-Dinescu et al. (2011) analyzed some thousands of red clump stars and obtained steeper gradients of $17.4 \pm 2.5 \text{ km s}^{-1} \text{ kpc}^{-1}$ in σ_u , $17.1 \pm 5.0 \text{ km s}^{-1} \text{ kpc}^{-1}$ in σ_v , and $5.0 \pm 2.4 \text{ km s}^{-1} \text{ kpc}^{-1}$ in σ_w from a subset of stars at $|Z| = 0.7\text{--}2.5 \text{ kpc}$, and they argue that this apparent disagreement could be attributed to the fact that Girard et al. averaged the steeper gradient at low $|Z|$ over a wider range of $|Z| = 1\text{--}4 \text{ kpc}$.

The kinematical properties of the Galaxy were studied extensively by Bond et al. (2010) using the SDSS DR7 sample of $\sim 10^5$ blue main-sequence stars at high latitudes ($|b| > 20^\circ$). Their analysis of 10^4 disk stars at $|Z| = 1\text{--}4 \text{ kpc}$ with $[\text{Fe}/\text{H}] > -0.9$ showed that $\sigma_u, \sigma_v, \sigma_w$, and $\langle V \rangle$ are proportional to $|Z|^{1.25\text{--}2}$ with their shallower gradients at low $|Z|$ becoming steeper toward larger $|Z|$. If their linear Z gradients are averaged over the range of $|Z|$ considered, they correspond roughly to $\sim 10 \text{ km s}^{-1} \text{ kpc}^{-1}$ for the velocity dispersion gradients and $\sim -30 \text{ km s}^{-1} \text{ kpc}^{-1}$ for the rotation gradient.

Although the detailed kinematical Z profile of the thick disk remains to be determined, it is certain that the disk-like kinematics changes rather smoothly up to $|Z| \sim 4 \text{ kpc}$ but shows no smooth transition to the halo kinematics beyond that height. Thus, the thick-disk component is not isothermal and is dynamically distinct from the halo. This conclusion is based on a comparison of kinematical properties binned in different ranges of $|Z|$; another insightful method is to sort the kinematics by metallicity.

The metal abundances of stars in the likely thick-disk range of $|Z| \sim 1\text{--}4 \text{ kpc}$ are distributed broadly around a mean value of $[\text{Fe}/\text{H}] \sim -0.8 \text{ dex}$, and several of them on the metal-poor end were earlier found to have disk-like rotation or higher angular momentum relative to the halo (e.g., Norris 1986). Whether they are identified as thick-disk stars at the metal-poor tail or simply as rare peculiars can in principle be determined by statistical analysis of rotational velocities binned in different ranges of $[\text{Fe}/\text{H}]$.

A rotational velocity distribution for stars at $|Z| \sim 1\text{--}4 \text{ kpc}$ shows some asymmetrical features whose shape depends on the metallicity. In fact, a double-peak feature appears toward lower metallicity, and the distribution with this feature is well fitted by a sum

of at least two Gaussian distributions $N(V) = n_d(2\pi\sigma_d^2)^{-1/2}\exp(-1/2(V - V_d)^2/\sigma_d^2) + n_h(2\pi\sigma_h^2)^{-1/2}\exp(-1/2(V - V_h)^2/\sigma_h^2)$, representing the thick-disk and halo components, respectively. The parameters determined by fits to the kinematical data are typically $(V_d, \sigma_d) \approx (180, 30) \text{ km s}^{-1}$, and $(V_h, \sigma_h) \approx (40, 90) \text{ km s}^{-1}$ (e.g., Bond et al. 2010; Spagna et al. 2010).

If metal-poor stars with thick-disk kinematics are rare peculiars, the naïve expectation is that the relative normalization n_d/n_h decreases steadily as the metallicity decreases; that is, the thick disk dominating the halo becomes less and less visible ($n_d/n_h \ll 1$) toward lower metallicity.

In contrast, however, a double-peak feature ($n_d/n_h \sim 1$) persists below $[\text{Fe}/\text{H}] \sim -1$ (e.g., Spagna et al. 2010) and even down to ~ -2 (e.g., Beers and Sommer-Larsen 1995; Carollo et al. 2010; Chiba and Beers 2000). This indicates that a sufficient number of metal-poor stars with disk-like kinematics are well separated and manifest themselves against the halo. They would more likely be associated with the metal-poor tail of the thick disk component, which may support earlier claims for a so-called metal-weak thick disk.

5.2 Metallicity Distribution Constraints

The metallicity distribution of stars at $|Z| \sim 1\text{--}4 \text{ kpc}$ also shows some asymmetrical features, which could be fitted by two Gaussian distributions, one with a mean $[\text{Fe}/\text{H}] \approx -0.8$ dex and a dispersion $\sigma \approx 0.2$ dex for the disk and another with $[\text{Fe}/\text{H}] \approx -1.5$ dex and $\sigma \approx 0.3$ dex for the halo (e.g., Ivezić et al. 2008). Further decomposition of the disk is not obvious, but apparently the thick disk dominates at such heights well beyond the scale height of the thin disk.

The metal-poor part of the Gaussian distribution for the potential thick disk overlaps considerably with the metal-rich part of that for the halo, and the overlap covers a range from $[\text{Fe}/\text{H}] \sim -1$ to -1.5 . Because stars in this range were born from gas that had experienced a similar level of chemical enrichment, a chemical continuity should have occurred in the transition between the thick disk and the halo, although the two components are *dynamically* discrete.

Given that the metal-poor limit for the thick disk is as low as $[\text{Fe}/\text{H}] \sim -2$, it is also intriguing to ask how far the metal-rich limit is extended. This problem has been discussed using high-resolution spectroscopic survey observations of nearby stars. Their kinematical properties are found to bifurcate into two components representing the thin and thick disks, respectively, and a considerable fraction of sample stars with near-solar metal abundances could be classified as belonging to the thick disk according to their kinematics (e.g., Bensby et al. 2003, 2005; Mishenina et al. 2004; Reddy et al. 2006). Thus, Bensby et al. (2007) explicitly emphasized that the metal-rich limit is as high as the solar metallicity or even beyond it, perhaps reaching $[\text{Fe}/\text{H}] \sim +0.2$. If this is really the case, the metal-rich part of the metallicity distribution for the thick disk must overlap fully with that for the thin disk.

In summary, an apparent chemical overlap appears at $[\text{Fe}/\text{H}] \sim -1$ to -1.5 between the thick disk and the halo and also at $[\text{Fe}/\text{H}] > -0.6$ between the thick and thin disks. However, the overlap does not necessarily indicate that such stars having similar metal abundances share a chemical origin. This is because the same value of the stellar metal abundance results from different histories of chemical enrichment depending on the age of the component and the timescale of star formation in it. Thus, age dating of the thick disk is at least useful in deducing its chemical evolution as well as its relationship to the other components, the thin disk and halo.

5.3 Age Constraints

In MFA of star counts, the age of the thick disk typical of $[\text{Fe}/\text{H}] \sim -0.8$ is estimated as the age of template globular cluster 47 Tuc, which is best fitted to the thick-disk component in the star count data (e.g., Gilmore and Reid 1983; Yoshii et al. 1987). In general, absolute age determination based on isochrone fitting still differs from author to author because of uncertainties in stellar models, but the use of the same stellar model yields a rather reliable estimate of relative ages among stars either in clusters or in fields. By this reasoning, the thick-disk cluster 47 Tuc is estimated to be about 3 Gyr older than the oldest stars in the thin disk (Liu and Chaboyer 2000), and conversely, it is coeval with mildly metal-poor halo clusters such as NGC 6652 and 1851 but 2–3 Gyr younger than more metal-poor halo clusters such as M3 and M92 (Chaboyer et al. 2000; VandenBerg et al. 2000).

Stellar ages can be estimated more directly by applying the isochrones to individual field stars in the sample. Although these ages may suffer from systematic uncertainties, their plots against metallicity based on high-resolution spectroscopic observations show an age–metallicity relation that reveals the general trend of chemical enrichment on a statistical basis. The rather dramatic result is that the thick disk, starting at an age of ~ 12 Gyr, may have become chemically enriched to solar and higher metallicities in a quite short period of 1–2 Gyr, followed by the thin disk starting about 8–9 Gyr ago; the thin disk then gradually reached solar and higher metallicities in an extended period of 5 Gyr or so (Bensby et al. 2007; Ramírez et al. 2007; Reddy et al. 2006). Accordingly, the results of age dating indicate that the Galaxy must have evolved in good sequential order beginning with the halo and thick disk and then proceeding to the thin disk.

5.4 Elemental Abundance Constraints

Chemical enrichment is driven by supernovae explosions, which release synthesized metals into the interstellar medium from which new stars are born. The metallicity at a certain elapsed time t since the beginning of star formation is the integral of the value over the past up to t , and a large or small rate of star formation directly causes rapid or slow enrichment, respectively.

The abundances of heavy elements ejected through supernovae explosions and the lifetimes of their progenitor stars both depend on the stellar mass, and these mass dependencies are responsible for differential enrichment in heavy elements. The elemental abundance pattern at t is obtained by averaging the abundances of heavy elements from explosions of progenitors having various masses m whose lifetimes t_m are shorter than the timescale of star formation, which is roughly comparable to t . To the degree that this instantaneous recycling condition of $t_m \ll t$ is met for any m of the progenitors, the averaged pattern over m is independent of t as well as of the metallicity, which is a function of t . On the other hand, the abundance ratio between any two heavy elements is sensitive to the initial stellar mass function (IMF), which produces a mass-dependent weight in the above averaging. In other words, the abundance ratio can then be used as a diagnostic test of the slope of the massive part of the IMF.

Explosive nucleosynthesis calculations show that type II supernovae (SNe II) from massive stars ($m > 8 M_\odot$) release many alpha and iron-peak elements, and more massive stars tend to yield a larger ratio of alpha to iron abundances. In addition, type Ia supernovae (SNe Ia) from intermediate-mass stars ($m \sim 3\text{--}5 M_\odot$) release a large excess of iron relative to alpha elements. Because SN Ia progenitors are considered to evolve on much longer timescales than those

of SN II, the iron source switches from SNe II to SNe Ia, which imprints a break in the alpha-to-iron ratio at a certain metallicity corresponding to the time at which a significant number of SN Ia progenitors start to explode.

Although the evolutionary timescales of SN II progenitors of single stars or their lifetimes $t_m \sim 10^{6-7}$ years are well established, those of SNe Ia thought to arise from binary systems are difficult to evaluate using theoretical arguments because of many complexities such as the primary-to-secondary mass ratio, initial separation, and accretion rate by mass flow. Instead, the evolutionary timescale of SNe Ia has been estimated from the trend in $[O/Fe]$ using a major α element of oxygen for nearby disk stars based on data published by Edvardsson et al. (1993) and references therein.

Yoshii et al. (1996) obtained a timescale of $t_{\text{SNIa}} \sim 1-2$ Gyr by fitting standard infall models of chemical evolution to data showing a break in $[O/Fe]$ at $[Fe/H] \sim -1$ for the solar-neighborhood stars presumably belonging to the thin disk. Moreover, the constant level of $[O/Fe] \sim +0.4$ at lower metallicities before the break was found to be consistent with the usual Salpeter-like IMF (Tsujiimoto et al. 1995). If this timescale of t_{SNIa} could apply anywhere beyond the solar neighborhood, the break in $[\alpha/Fe]$ is expected to occur at different metallicities for different degrees of chemical enrichment from the thin disk.

Whether this nucleosynthetic signature of SNe Ia is also seen for the thick disk has been examined by high-resolution spectroscopic observations of nearby stars classified as members of either the thin disk or the thick disk using their kinematical properties (Bensby et al. 2003, 2004; Feltzing et al. 2003; Fuhrmann 1998; Mishenina et al. 2004; Prochaska et al. 2000; Ramírez et al. 2007; Reddy et al. 2006). All these observations indicated that the value of $[\alpha/Fe]$ at a given $[Fe/H]$ for thick-disk stars is enhanced over that for thin-disk stars in the observed range of $[Fe/H]$.

In fact, the thick disk sustains a constant level of $[\alpha/Fe] \sim +0.4$ up to a higher metallicity $[Fe/H] \sim -0.4$, in contrast to the thin disk, where $[\alpha/Fe]$ starts to decrease as $[Fe/H]$ increases toward the solar metallicity. This finding of the break in $[\alpha/Fe]$ clearly confirms the explosions of SNe Ia that enriched the gas. The immediate consequences are two-fold: the thick disk has experienced rapid chemical enrichment to reach $[Fe/H] \sim -0.4$ until the onset of SNe Ia at $t \sim t_{\text{SNIa}} (\sim 1-2$ Gyr) and the thick disk has retained sufficient gas inside to allow new stars to form thereafter. Thus, it is suggested that the thick disk is older than the thin disk, but this is more explicitly demonstrated by the age-metallicity relations reported for the two disks (Bensby et al. 2007; Ramírez et al. 2007; Reddy et al. 2006). In terms of $[\alpha/Fe]$ and age, these two disks are discrete, and the possibility that the thin disk expanded into the thick one would be no more than an interesting speculation.

6 Formation and Evolution of Thick Disk

6.1 Summary of Observational Constraints of the Thick Disk

Many observations of the thick disk have *not* diverged but fortunately suggest a rather convergent view of its structure, kinematics, and chemistry. The essential points revealed for the thick disk are itemized as follows:

1. The thick disk is, like the thin disk, represented by a double exponential density law in the Z direction perpendicular to the disk plane and also in the cylindrical radial direction away

- from the Galactic center. The thick disk has the scale height $h_Z \sim 750$ pc and length $h_R \sim 3.5$ kpc, with a normalization of about 10% of the thin-disk density near the sun. These scale sizes are larger than those of the thin disk, that is, $h_{Z,\text{thick}}/h_{Z,\text{thin}} \sim 2$ and $h_{R,\text{thick}}/h_{R,\text{thin}} \sim 1.2$. A good measure of the mass ratio yields $M_{\text{thick}}/M_{\text{thin}} \equiv (\rho_0 h_R^2 h_Z)_{\text{thick}}/(\rho_0 h_R^2 h_Z)_{\text{thin}} \sim 0.3$.
2. The thick disk exists not only in the Galaxy but also in external disk galaxies. The surface brightness distribution along the minor and major axes of edge-on disk galaxies is fitted by thick and thin disks, each of which is represented by a double exponential law. The scale height and length of the two disks increase with increasing galactic mass, but the scale ratios between the two disks stay almost constant with some dispersion, and their values agree with those of the Galaxy.
 3. The thick disk has a typical W -velocity dispersion $\sigma_w \sim 40$ km s⁻¹ at $|Z| \sim 1\text{--}3$ kpc, but it has a clear vertical gradient and also gradients in the other two directions. Thus, the thick disk is not isothermal in the Z direction, invalidating the use of the sech² function for the density distribution perpendicular to the disk plane. To be consistent with the observed exponential density law, the σ_w -dependent normalization must be introduced in a strictly fine-tuned manner.
 4. The thick disk has a vertical gradient of mean rotational velocity $dV/d|Z|$, the value of which remains unclear between two clusters of estimates either shallower or steeper than -20 km s⁻¹ kpc⁻¹. Plausible thick-disk stars in the range of $|Z| = 1\text{--}3$ kpc lag the thin-disk rotation by ~ 40 km s⁻¹, so the ratio of specific angular momenta is roughly $j_{\text{thick}}/j_{\text{thin}} \equiv (h_R V)_{\text{thick}}/(h_R V)_{\text{thin}} \sim 1$. Accordingly, the ratio of total angular momenta is $J_{\text{thick}}/J_{\text{thin}} \sim M_{\text{thick}}/M_{\text{thin}} \sim 0.3$.
 5. The thick disk meets the condition of rotational equilibrium such that $U \ll \sigma_u$, $W \ll \sigma_w$, and $V \gg \sigma_v$ and has a radially elongated velocity ellipsoid of $\sigma_u/\sigma_v \sim \sigma_u/\sigma_w \sim 1.5\text{--}2$, similar to the thin disk, which suggests that a third isolating integral was maintained in the transition between two disks. The ratio of random kinetic energies per unit mass is roughly $(1/2 \sigma^2)_{\text{thick}}/(1/2 \sigma^2)_{\text{thin}} \sim 5$ in each direction, so this change between the two disks occurs without any preferred direction.
 6. The thick disk exhibits a fairly large scatter in metallicity around a mean $[\text{Fe}/\text{H}] \sim -0.8$ dex, and its Z gradient is shallower than $d[\text{Fe}/\text{H}]/d|Z| = -0.1$ dex kpc⁻¹, even allowing the possibility of no gradient at all. This is to be compared with the properties of the thin disk, such as a smaller scatter of $-0.5 \leq [\text{Fe}/\text{H}] < +0.3$ and a steeper gradient of $d[\text{Fe}/\text{H}]/d|Z| = -0.5$ dex kpc⁻¹. Therefore, although the thin disk has a chemically stratified Z structure, the thick disk shows no such stratification.
 7. The thick disk has almost no metallicity gradient $d[\text{Fe}/\text{H}]/dR$ in the radial direction. Together with other aspects of the thick disk, such as the small vertical gradient of metallicity and the extreme metallicity tails reaching as far as $[\text{Fe}/\text{H}] \sim -2$ dex and oppositely $+0.2$ dex, the thick disk is more like a chemical mixture in which a wide variety of metallicities are scattered throughout the entire thick-disk region. This contrasts sharply with the chemically stratified Z structure of the thin disk.
 8. The thick disk shows a signature of SN Ia explosions at $[\text{Fe}/\text{H}] \sim -0.4$ dex, compared to $[\text{Fe}/\text{H}] \sim -1$ for the thin disk, where the constant level of $[\alpha/\text{Fe}] \sim +0.4$ starts to decrease toward solar metallicities. Because this break occurs at a time corresponding to the evolutionary timescale of $t_{\text{SNIa}} \sim 1\text{--}2$ Gyr, the difference in the break metallicity indicates that the thick disk has undergone more rapid chemical enrichment, making its metallicity higher than that of the thin disk.

9. The thick disk has a typical age of 10 Gyr, which is intermediate between halo globular clusters, which are a few Gyr older, and thin-disk stars, which are a few Gyr younger. Isochrone fits to individual stars belonging to the thick and thin disks according to their kinematics revealed that chemical enrichment began in the thick disk some 12 Gyr ago and quite rapidly reached solar metallicities in 1–2 Gyr, whereas it began in the thin disk about 8–9 Gyr ago and reached solar metallicities in about 5 Gyr.

6.2 Possible Scenarios

A variety of formation scenarios for the thick disk have been proposed, but they belong to essentially two categories: (1) stars that formed outside moved to form the thick disk and (2) stars that formed in situ remained to form the thick disk.

The former category has three subclasses: accretion of stars from disrupted satellites (e.g., Abadi et al. 2003), heating of stars in the preexisting thin disk by external impacts (e.g., Villalobos and Helmi 2008), and radial migration of stars through resonant scattering by spiral arms (e.g., Schönrich and Binney 2009). Each of them may indeed explain some of observational properties #1 through #9, but not all of them. The accretion scenario must invoke at least some mechanism for the origin of the double exponential law for the thick and thin disks (#1, 3) and the constant scale ratios between the two disks (#2). The need for a favorable inclination, direction, velocity, and mass of accreting satellites is equivalent to requiring the initial conditions to be strictly fine-tuned for each disk galaxy, which requires further explanation. The heating of the thin disk, which suffers from the same problem of fine-tuning, evidently conflicts with the α element overabundances of the thick disk compared to the thin disk (#8) and with the greater age of the thick disk compared to the thin disk (#9). The migration of thin-disk stars through resonant scattering spreads the radial distribution of stars by only a few kiloparsecs and cannot form the entire thick disk, which is radially more extended than the thin disk (#1, 3).

The latter category explicitly includes star formation from the turbulent gas in a slowly rotating component formed after halo collapse (e.g., Bournaud et al. 2009) or in the late-phase gas-rich mergers motivated by cosmological simulations (e.g., Brook et al. 2004). In either case, it is common for gas clumps to collide with other clumps to form cool, dense regions from which stars are born. The apparent initial difference tends to vanish through dissipational contraction and spin-up, the timescale of which depends on the spatial separation among the star-forming sites considered.

Rapid enrichment in the thick disk is driven by a large rate of star formation, necessarily causing a large rate of supernovae explosions, which release an enormous amount of kinetic energy into the interstellar gas. Thus, the problem of survival arises because the energy input from SNe may expel gas from the star-forming sites and halt the formation of new stars before chemical enrichment in the thick disk reaches the solar metallicity. One extreme way to escape from this problem is to assume that the star-forming sites are scattered in space with sufficiently close separation among them. In this case, the gas lost from one site is incorporated into neighboring sites, and new star formation continues there. Thus, the chemical evolution of the thick disk, if averaged over the entire region, is described by a competition between gas consumption in star formation and the net loss of gas that falls onto the disk plane and eventually forms the thin disk. The other extreme is to assume that star-forming sites are as massive as dwarf galaxies and can gravitationally confine the gas against the energy input from SNe. In this case,

self-enrichment proceeds within individual massive sites, many of which would be disrupted at their passage across the disk plane and eventually form the thick disk.

The true situation may lie between these opposite extremes, but any scenario of in situ formation of the thick disk between the halo and the thin disk implies that gas clumps in the rotationally supported, extended protodisk have larger random velocities than those of the thin disk, and therefore have a greater chance of encounters, which leads to a larger star formation rate as well as a larger rate of angular momentum exchange. Accordingly, this scenario explains the more rapid chemical enrichment of the thick disk (#8) compared to the thin disk and the thick disk's intermediate kinematics (#3, 4, 5), chemistry (#6, 7), and age (#9) between those of the halo and the thin disk.

Unlike the usual assumption regarding chemical evolution, however, the gas in the star-forming sites is neither homogeneous nor well mixed. The rate of star formation, which is presumably proportional to some power of the gas density (Schmidt 1959), is quite different from site to site. In the final state of chemical enrichment, sites having different metallicities are scattered over the entire thick-disk region, and an ensemble of stars formed in these sites likely shows a large scatter of metallicity around the mean without any substantial gradient (#6, 7). Thus, dissipational contraction, which yields the observed kinematical gradient, does not necessarily yield a metallicity gradient in the thick disk.

The remaining issue to be explained is the universality of the double exponential structure of the thick and thin disks (#1, 2, 3). A number of authors simulated the collapse and evolution of a proto-Galaxy (e.g., Burkert et al. 1992) and cosmological Λ -cold dark matter (Λ -CDM) simulations of Galaxy formation (e.g., Brook et al. 2004), taking into account many physical processes thought to occur in a multiphase interstellar medium. In such simulations, the double exponential structure of stellar disks appears in the end of evolution and the reason for this must be explored, as described in the next section.

6.3 Origin of Double Exponential Stellar Disk

Double exponential luminosity profiles in the vertical and radial directions in disks, as well as a flat rotation curve, are two apparently universal features of disk galaxies, independent of whether they formed in isolation or in a cluster environment. The exponential luminosity profile is usually interpreted as an exponential mass distribution, where the mass-to-luminosity ratio is assumed to be constant. These structural similarities may have emerged in different environments due to some internal regularities that dominate external disturbances such as interactions and mergers.

Among the many physical processes involved, star formation distinguishes galaxy simulations from merely hydrodynamical computations. The rate of star formation, although it heavily influences the end result of galaxy simulations, is not understood from first principles because of our limited understanding of the fundamental physics of star formation. Ideally, simulations with various assumptions of star formation are repeated, but a large parameter space cannot be fully explored in detail because of computational limitations. One useful pathway to progress is to search for an underlying mechanism, beyond the ad hoc specification of star formation in simulations, that accounts for at least some of the well-established observations of real galaxies. This section describes several trials along this line and shows that the hydrodynamical equations associated with some simple hypotheses for star formation yield a remarkable solution that naturally produces the basic structural features of galaxies in their final stages of evolution.


6.3.1 Vertical Exponentiality

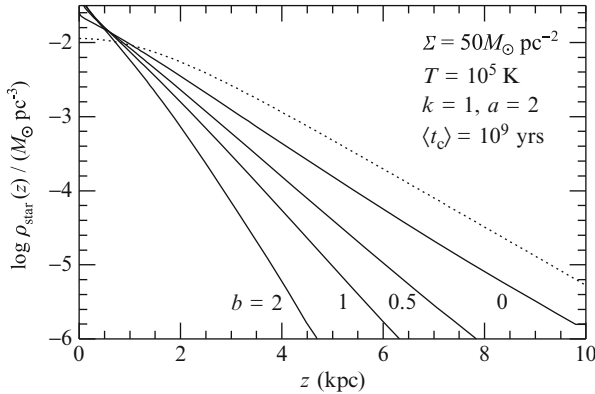
There is strong evidence from optical observations that galactic disks have a universal luminosity profile perpendicular to the disk plane or, equivalently, a universal mass distribution in the Z direction assuming a constant mass-to-luminosity ratio (Burstein 1979; van der Kruit and Searle 1981). Near-infrared observations of edge-on spiral galaxies have uncovered an excess over an isothermal distribution at small $|Z|$ that was obscured in previous optical observations (Aoki et al. 1991; de Grijs et al. 1997; Wainscoat et al. 1989). An exponential Z distribution can be constructed by adding up several stellar disk components with different vertical velocity dispersions, but this is possible only if the contribution from populations with larger velocity dispersions is fine-tuned to dominate progressively at larger distances from the disk plane. The underlying mechanism for this vertical fine-tuning is identified with gravitational settling of the protodisk, which naturally produces the exponential Z distribution of stars as long as the timescale of star formation t_s is comparable to that of gas cooling t_c , which induces dissipational contraction of the protodisk (Burkert and Yoshii 1996).

A large fraction of the halo gas collapses, or else gas-rich mergers occur, to grow an initially extended protodisk, in which gas clumps move randomly and some of them collide with others. Through inelastic collisions, the kinetic energy of random motion is thermalized to excite atoms and then lost by their radiative deexcitation. At the same time, this cooling of the gas plays a decisive role in efficient star formation, as it provides the required environment in which density fluctuations in the gas can become gravitationally unstable on stellar mass scales.

The protodisk, which may not be in dynamical equilibrium initially, is expected to eventually achieve a quasi-equilibrium state in the Z direction because the increase in the energy input from supernova explosions compensates for the energy dissipation by cooling. Thereafter, the rate of star formation is adjusted to balance the rate of energy dissipation by means of a self-regulated star formation mechanism (Cox 1983), and the protodisk contracts by dissipating its kinetic energy on a timescale t_c while forming stars on a timescale t_s .

The evolutionary behavior of such a protodisk is determined by the ratio $k = t_s/t_c$, which should be of order unity according to the hypothesis that stars are formed on t_s comparable to t_c and remain for most of the time at the Z distance where they initially formed. In the extreme case of $k \ll 1$, the initial gas distribution is frozen in the final stellar disk; in the opposite extreme of $k \gg 1$, the gas contracts without limit, producing a highly concentrated stellar disk on the plane. Only the case of $k \sim 1$ between these two extremes yields an exponential stellar Z distribution in the final configuration.

The cooling rate of the protodisk is a function of the local gas density and temperature and therefore can be approximated as $\Lambda \propto \rho^a T^b$. Typical values are $a = 2$ and, for ionized hydrogen gas, $b = +0.5$ (free-free transition) or -0.5 (free-bound transition), but the combination (a, b) can also be left as a free parameter.  *Figure 8-13* shows a sequence of models with changing b , while the other parameters are fixed with $k = 1$. The final stellar Z distribution is perfectly exponential, spanning about five orders in density, and its exponential scale height h_Z becomes smaller for larger b and generally differs from the initial scale height of the protodisk. This remarkable result holds independent of the prescribed (a, b) and initial physical state of the gas, provided that k is of order unity and the protodisk reaches quasi-equilibrium before efficient formation of disk stars. Note that a larger initial column density Σ yields a larger-scale height, which is consistent with the observed trend that massive galaxies have stellar disks with larger-scale heights (Collins et al. 2011; Yoachim and Dalcanton 2006).



■ Fig. 8-13

Final stellar Z distributions for different values of the power index b for the cooling rate $\Lambda \propto \rho^a T^b$ with $a = 2$ and other parameters fixed as shown in the panel. *Thick lines* indicate the final stellar Z distributions arising from the same initial distribution of the protodisk gas (*dotted line*). Note that a larger initial column density Σ produces a larger final scale height

The vertical velocity dispersions of $\sigma_w \sim 40 \text{ km s}^{-1}$ for the thick disk and $\sim 10 \text{ km s}^{-1}$ for the thin disk, if thermalized, correspond to temperatures of $T \sim 10^5 \text{ K}$ and 10^4 K , respectively. When gas clumps collide with each other at this velocity, the gas is heated to the corresponding temperature and then cooled by the radiative deexcitation operating at these temperatures. The cooling rate at $T \sim 10^5 \text{ K}$ is dominated by the hydrogen free-bound transition with $b = -0.5$, and that at $T \sim 10^4 \text{ K}$ is dominated by the hydrogen $2p, s \rightarrow 1s$ transition with $b \approx 2-3$ (e.g., Boehringer and Hensler 1989). The inclusion of helium ($n_{\text{He}}/n_{\text{H}} = 0.1$) and a typical thick-disk metallicity ($z/z_{\odot} = 0.2$) enhances the cooling rate at $T \sim 10^5 \text{ K}$ and weakens its temperature dependence there.

Accordingly, dissipational contraction of the protodisk starts with $b = -0.5$ or 0 , depending on how much metallicity is included in the gas. Assuming $k \sim 1$, the stellar disk becomes exponential, having a scale height $h_Z(b \leq 0)$ for the value of b specified in the above range. As the contraction proceeds, the dominant cooling process switches to that with $b \approx 2-3$, and the resulting stellar disk after this epoch has a distinctly smaller-scale height $h_Z(b \geq 2)$. Inspection of ● Fig. 8-13 indicates that $h_Z(b \leq 0)$ is at least two times larger than $h_Z(b \geq 2)$, which agrees well with the observations if these two disks are identified as the thick and thin disks, respectively. The hypothesis of $k \sim 1$ and atomic physics therefore combine to provide a firm theoretical reason for the existence of two scale heights ($h_{Z,\text{thick}}, h_{Z,\text{thin}}$) in a galactic disk.

6.3.2 Radial Exponentiality

The exponential luminosity profile in the radial direction, or the radial distribution of the exponential stellar disk, was known long before the vertical profile was recognized (Freeman 1970). The underlying mechanism is identified with viscous evolution of the star-forming gas disk,

which produces a nearly exponential stellar disk in the radial direction, irrespective of the initial conditions, as long as the timescale of star formation t_s is comparable to that of the kinetic viscosity t_v (Lin and Pringle 1987b; Yoshii and Sommer-Larsen 1989).

In the early stage, an initially extended protodisk consisting of gas clumps sooner or later reaches a state of centrifugal equilibrium in differential rotation. The protodisk is then expected to evolve by exchanging angular momentum on t_v and simultaneously forming stars on t_s . There are primarily two sources of kinetic viscosity that radially redistribute the angular momentum in a differentially rotating disk: inelastic collisions of gas clumps (Silk and Norman 1981) and instabilities due to self-gravity (Toomre 1964), both of which enhance the density fluctuations leading to star formation. Thus, the ratio $k = t_s/t_v$ should be of order unity because these processes originate from the same mechanism. A useful hypothesis here is therefore $k \sim 1$, similar to the case in the vertical direction.

The viscosity coefficient is prescribed as a function of the local surface gas density and angular velocity of rotation in the disk; therefore, it can be approximated as $\nu \propto \Sigma^a \Omega^{-b}$. Assuming a flat rotation curve, a combination of $(a, b) = (1, 2)$ is obtained for the collisions of gas clumps (Olivier et al. 1991), and $(2, 3)$ is obtained for the self-gravitating disk (Lin and Pringle 1987a), but it may also be reasonable to leave the viscosity unspecified because of uncertainties in its prescription.

Extensive calculations of viscous evolution with $k \sim 1$ show that the distribution of the surface density of stars becomes nearly exponential over a wide range of disk radius up to about six scale lengths. The end product depends little on the initial conditions and viscosity prescription (a, b) , provided that k is of order unity and the resulting scale length of the stellar disk is about one fourth of the outer radius of the initial gas disk (Lin and Pringle 1987b; Olivier et al. 1991; Yoshii and Sommer-Larsen 1989). This indicates that a larger initial gas disk produces a stellar disk with a larger-scale length, which is consistent with the observed trend of external disk galaxies (Collins et al. 2011; Yoachim and Dalcanton 2006).

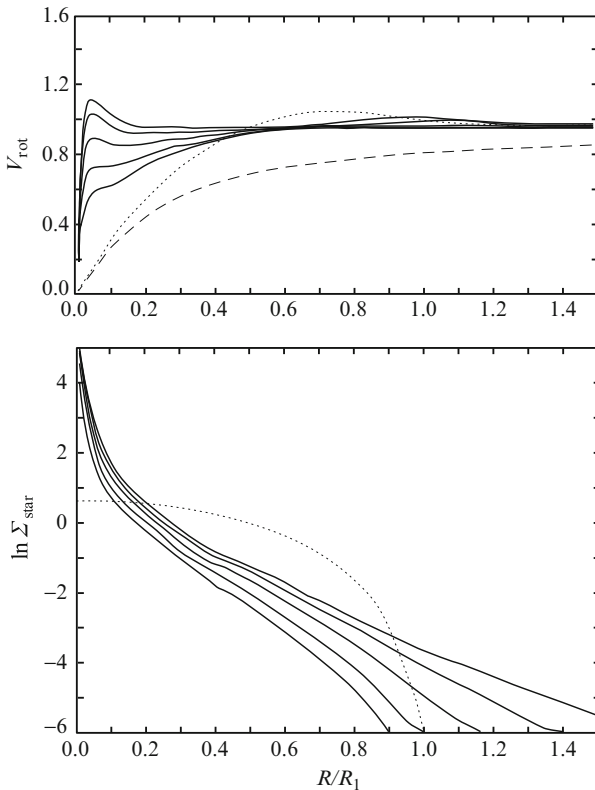
In actuality, much of the gas in an initially extended protodisk moves inward by the outward transfer of angular momentum and falls spirally onto the disk plane at smaller radii by energy dissipation. After the exponential stellar thick disk forms, viscous evolution of the thin disk begins in a more centrally accumulated distribution of the gas. Because the gas clumps that form the thin disk move in nearly circular orbits with smaller random velocities, the exchange of angular momentum across circular annuli is less efficient, causing a more elongated evolution of the thin disk. Hence, the resulting stellar thin disk becomes exponential on a longer timescale with a smaller-scale length compared to the thick disk.

A limit on the radial contraction of the gas disk may be set by the ratio $r_{\text{ap}}/R_{\text{circ}}$, namely, the apogalactic radius for gas clumps of the thick disk relative to the circular radius for those of the thin disk with the same angular momentum value. A rough measure of $r_{\text{ap}}/R_{\text{circ}}$ from the typical orbital eccentricity of thick-disk stars $e \sim 0.2\text{--}0.3$ (Carollo et al. 2010; Casetti-Dinescu et al. 2011; Dierickx et al. 2010; Wilson et al. 2011) gives $h_{R,\text{thick}}/h_{R,\text{thin}} \sim r_{\text{ap}}/R_{\text{circ}} \sim 1.2$, which is more or less consistent with the recent observational trend. As a result, the hypothesis $k \sim 1$ provides a firm theoretical reason for the existence of two scale lengths ($h_{R,\text{thick}}, h_{R,\text{thin}}$) for a galactic disk.

Additionally, viscous evolution involves a far more promising solution to the origin of the flat rotation curve all the way from the inner region dominated by the self-gravity of the disk, to the outer region governed by an isothermal, spherical dark halo surrounding the disk. The transition from the inner region to the outer region mysteriously imprints no features in the rotation curve. The existence of flat, featureless rotation curves indicates that the peak rotational velocities arising from these different components of the disk and halo must be the same

(Bahcall and Casertano 1985). This so-called disk–halo conspiracy is explained by viscous evolution with $k \sim 1$, which simultaneously produces an exponential stellar disk in the radial direction and a flat rotation curve, irrespective of the initial conditions (Saio and Yoshii 1990).

► *Figure 8-14* shows the evolution of the rotation curve and the surface densities of stars. As evolution proceeds, the rotation curve in the outer part of the disk becomes nearly constant at a velocity determined by the gravity of the dark halo, and the distribution of the surface density of stars becomes nearly exponential over a wide range of disk radius. Both the central shape of the rotation curve and the central concentration of stars depend on the ratio $k = t_s/t_v$ and also on the specific angular momentum $j = J/M$ given to the initial disk. More centrally concentrated structures arise from models with larger k or lower j , allowing for later growth of the central bulge in the course of thick to thin disk formation even after the halo gas has collapsed. Because the timescale t_v in the central region is smaller, causing more rapid chemical enrichment, such later growth may partially explain the observed similarity of elemental abundances between thick-disk and bulge stars (Meléndez et al. 2008).



■ Fig. 8-14

Viscous disk evolution of the rotational velocity (*upper panel*) and surface density of stars (*lower panel*) for a prescribed viscosity of $\nu \propto \Sigma^a \Omega^{-b}$ with $(a, b) = (2, 3)$. Evolutionary changes with time are shown by *solid lines* from bottom to top. *Dotted lines* represent the initial distributions. *Dashed line* is the rotational velocity arising only from the dark halo

The essence of the hypothesis $k \sim 1$ for the disk is that the timescale of local star formation is comparable to the timescale of material flow that eventually organizes the global structure of galaxies. For example, under the action of kinetic viscosity in a differentially rotating disk, the gas in the inner region moves toward the disk center, whereas that in the outer region moves in the opposite direction, away from the center (Lynden-Bell and Pringle 1974). In gravitational settling of the vertical disk, the gas undergoes one-way contraction onto the equatorial plane. Thus, regardless of which direction the gas moves, an exponential stellar distribution always results as long as $k = 1$. However, this hypothesis is not applicable to the halo.

6.4 Transition from Halo to Thick Disk

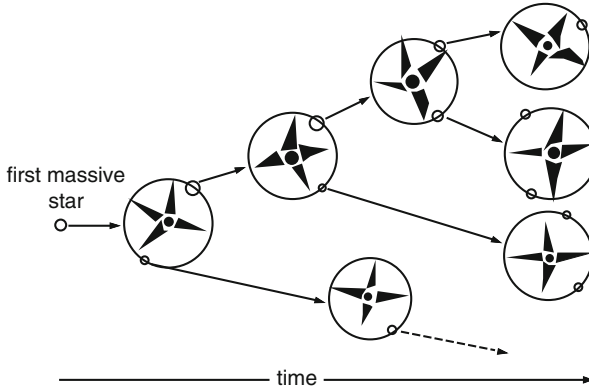
The evolutionary path from the halo to the thick disk is closely related to the question of whether the halo is broken into fragments (separate clouds) during its collapse or is assembled from CDM clouds and how halo stars are formed in these clouds. Although this problem has been explored by detailed N -body/smoothed-particle hydrodynamics (SPH) simulations, it is conjectured that halo stars form on a timescale t_s much longer than the timescale of gas infall t_f from the halo into the disk region. In the Galaxy, the observed metallicity distribution of halo stars requires $k = t_s/t_f \sim 10$ (Hartwick 1976; Ryan and Norris 1991), which necessarily implies that the properties of halo stars arising from $k \gg 1$ do not show a smooth transition to those of thick-disk stars from $k \sim 1$.

The validity of the assumption that halo stars are formed from chemically well-mixed gas has been challenged by the fact that the mixing length of heavy elements ejected from SNe is too small to achieve chemical homogeneity over the entire halo. Current estimates of the velocity dispersion and total mass of the stellar component in the Galactic halo indicate that consecutive star-forming processes are likely to have been confined in separate clouds having mass scales of $M \geq 10^{6-7} M_\odot$, which make up the baryonic halo. The first stars in a protogalaxy could be born in highly compressed layers behind shock waves caused by supersonic initial turbulence. The massive stars among them eventually explode as SNe, triggering the formation of stars in dense shells of mass M_{sh} . This SN-induced star formation proceeds for generations, as schematically illustrated in [Fig. 8-15](#), and terminates when SN remnants (SNRs) become unable to sweep up enough gas to form dense shells (Shigeyama and Tsujimoto 1998; Tsujimoto et al. 1999).

Given the mass fraction ε of stars formed in the dense shell of each SNR, the star formation rate is given by a recurrence relation,

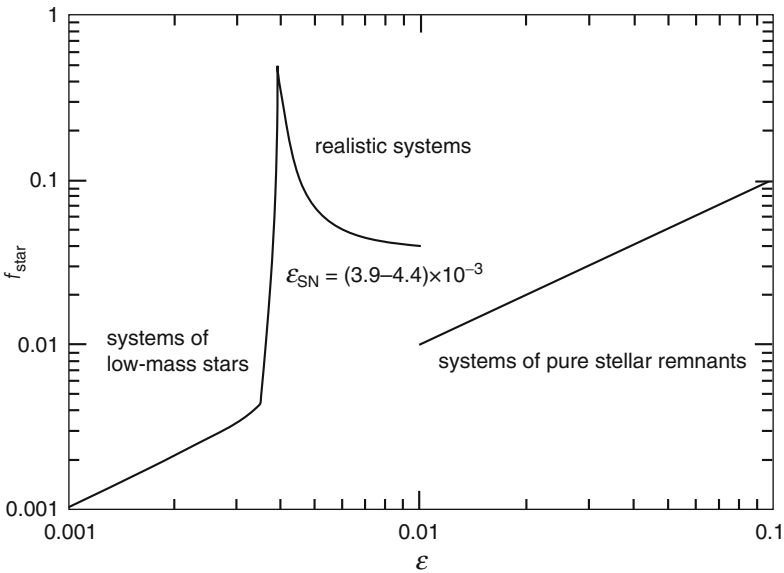
$$\dot{M}_s(t) = \int dm \varepsilon M_{\text{sh}}(t, m) [\phi(m)/m] \dot{M}_s(t - t_m),$$

where t_m is the lifetime of stars of mass m , $\phi(m)$ is the initial stellar mass function, and the integration runs over the mass range of SN II progenitors. The value of ε is chosen according to the chemical evolution that follows. As shown in [Fig. 8-16](#), this condition yields that ε must be strictly confined within a very narrow range of $\varepsilon_{\text{SN}} = (3.9-4.4) \times 10^{-3}$ for $M_{\text{sh}} = 6.5 \times 10^4 M_\odot$ (Shigeyama and Tsujimoto 1998), so the number of massive stars born from each SNR is nearly one, $N_{\text{SN}} = 1$. If $\varepsilon > \varepsilon_{\text{SN}}$ or $N_{\text{SN}} > 1$, the total number of SNe ever formed after the i -th generation diverges as N_{SN}^i , and star formation is soon terminated with little chemical enrichment. If $\varepsilon < \varepsilon_{\text{SN}}$ or $N_{\text{SN}} < 1$, then N_{SN}^i stays near or below unity, so the cloud turns into a system of low-mass stars with little or no chemical enrichment. Thus, only a value of ε in the above narrow range ultimately gives rise to chemically evolved stellar systems.



■ Fig. 8-15

Schematic illustration of SN-induced star formation for $N_{\text{SN}} = 2$, in which two massive stars are born from each SNR



■ Fig. 8-16

Sensitivity of the mass fraction of stars f_{star} formed in a cloud to the value of ϵ that determines how much of the mass $M_{\text{sh}} = 6.5 \times 10^4 M_{\odot}$ swept up by SNRs turns into new stars. Notice that only values of ϵ in a very narrow range give to chemically evolved systems

In particular, a chemical evolution model with $\epsilon \approx \epsilon_{\text{SN}}$ reproduces the observed $[\text{Fe}/\text{H}]$ distribution function of halo stars up to $[\text{Fe}/\text{H}] \sim -1$ with a mean of ~ -1.5 without appealing to the manipulation required by previous models by which the heavy-element yield from SNe II is effectively decreased. Only a few tenths of the initial gas in a cloud is converted to halo stars, and the rest is blown out of the cloud and subsequently falls into the protodisk. Stars that

formed in separate clouds are distributed over the entire halo region, and an ensemble of these stars likely shows a scatter of metallicity around the mean without any substantial gradient in space. The mechanism that explains most of the observed chemical properties of halo stars is that each SNR produces nearly one massive star. That is, a massive star, once formed in an SNR, starts to release ionizing radiation and a stellar wind, which inhibit effective formation of new stars in the same SNR.

The velocity dispersion for stars formed in separate clouds is much smaller than the velocity dispersion for an ensemble of clouds that make up the entire halo. Therefore, unless dynamical relaxation in the halo is complete, distinguishable clumps of halo stars are expected to remain in the angular momentum phase space (Helmi and White 1999). The existence of several such stellar clumps (Chiba and Beers 2000; Helmi et al. 1999) as well as stellar streams (Yanny et al. 2003) partially justifies the idea that star formation in a given cloud proceeds independently of other clouds, which have their own space motions and chemical histories.

Some of these stellar clumps, which originate in clouds with $\varepsilon \approx \varepsilon_{\text{SN}}$ and orbit near the disk plane, are tidally disrupted during passage through the disk plane. Because stars from these disrupted systems have halo-like metallicities and cover only the low-metal part of the thick disk, their accretion may at best account for its metal-weak tail (e.g., Wyse et al. 2006) but does not form the entire population of the thick disk, which shows a broad range of metallicities.

If ε is not in the narrow range of ε_{SN} , almost all of the gas in a cloud is blown out for $\varepsilon > \varepsilon_{\text{SN}}$, whereas it remains in the cloud for $\varepsilon < \varepsilon_{\text{SN}}$. A considerable amount of the gas either ejected from or remaining in clouds, which requires external impacts to resume star formation, is destined to fall onto the disk plane and spins up to form a rotationally supported, extended protodisk where thick-disk stars begin to form. Accordingly, halo stars formed before this spin-up rotate slowly, whereas thick-disk stars formed after the spin-up rotate rapidly, which yields the observed clear separation between the rotational velocities of halo and thick-disk stars.

In summary, the low efficiency of SN-induced star formation allows a sufficient amount of chemically unprocessed gas to fall and form an extended protodisk, where gas clumps, newly formed or surviving from the halo, collide with others to form thick-disk stars more efficiently than halo stars. Then, chemically processed gas expelled from the star-forming sites falls further onto the disk plane, where the usual Schmidt law is a good approximation for the formation of thin-disk stars. This sequence of gas infall from the halo to the thick disk and then from the thick disk to the thin disk is crucial to explaining the different means and dispersions of the metallicity distributions for these three stellar populations (Gilmore and Wyse 1986; Hartwick 1976; Yoshii 1984). Along with the evolutionary sequence, the timescale of gas infall, initially not in balance with SN-induced star formation during the evolution of the halo, becomes balanced with that of star formation in the thick disk. This recovery of balance predicts the kinematical discontinuity in the transition between the halo and the thick disk.

7 Future Directions

Since an extensive list of possible scenarios for thick-disk formation was published by Majewski (1993), the relevant stellar data have greatly expanded in both quantity and quality. Some of the scenarios have already become unlikely, and some have survived the resulting observational constraints. A comparison of existing data with the predictions of simple models, existing numerical simulations, and theoretical best guesses, seems to favor dissipational contraction of a disk-like gas component in which thick-disk stars are formed in situ with rapid chemical enrichment.

The distribution of the orbital eccentricity of thick-disk stars has recently been proposed as a useful filtering tool for evaluating various scenarios (Sales et al. 2009). By detailed comparisons of the predicted and observed eccentricity distributions, large spectroscopic surveys such as the Sloan Extension for Galactic Understanding and Exploration (SEGUE)/SDSS and Radial Velocity Experiment (RAVE) projects give a convergent result: in situ star formation scenarios, such as gas-rich mergers, are most likely (SEGUE: Carollo et al. 2010; Dierickx et al. 2010; RAVE: Casetti-Dinescu et al. 2011; Ruchti et al. 2011; Veltz et al. 2008; Wilson et al. 2011).

However, a definitive answer awaits not only further acquisition of data for stars at great distances from the sun, but also progress in high-resolution numerical simulations of galaxy formation. Although such data are the targets of the ongoing ground-based projects SEGUE and RAVE and a forthcoming satellite project Gaia, the current computational performance of N -body/SPH galaxy simulations is still far from resolving stellar mass scales.

In fact, “stars” in these simulations are associations of many stars or “star particles” of $\sim 10^{4-5} M_{\odot}$ at best, which are formed according to the usual prescription of star formation (Katz 1992). Any predictions of stellar properties based on these star particles are not straightforward and inherently involve theoretical uncertainty. In this regard, several star formation hypotheses, introduced in the preceding sections, are extremely useful in calculating, from simple formulations, the amount of kinetic energy, gas, and metals ejected from the star particles as well as the epoch of their ejection. With the use of these calculated quantities as theoretical “data” in chemical and dynamical simulations of galaxies, our understanding of the evolutionary path from the halo to the thick disk is expected to allow us to discriminate between various aspects of previously suggested scenarios not only for the halo but also for the thick disk.

While the *nature* of the Galactic thick disk will be understood in much more detail as large, precise stellar datasets become available in the near future, its *origin* will remain to be determined unless galaxy simulations reliably quantify possible scenarios of thick-disk formation.

Acknowledgments

YY acknowledges partial support from the Grant-in-Aids of Scientific Research (17104002) of the Ministry of Education, Science, Culture and Sports of Japan and thanks many colleagues for discussions and suggestions throughout the period of writing this chapter.

Cross-References

- [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- [Dynamics of Disks and Warps](#)
- [Galactic Distance Scales](#)
- [Globular Cluster Dynamical Evolution](#)
- [High-Velocity Clouds](#)
- [History of Dark Matter in Galaxies](#)
- [Interstellar PAHs and Dust](#)
- [Magnetic Fields in Galaxies](#)
- [Mass Distribution and Rotation Curve in the Galaxy](#)
- [The Infrared Galaxy](#)

References

- Abadi, M. G., Navarro, J. F., Steinmetz, M., & Eke, V. R. 2003, *ApJ*, 591, 499
- Alcobé, S., & Cubarsi, R. 2005, *A&A*, 442, 929
- Allende Prieto, C., Beers, T. C., Wilhelm, R., Newberg, H. J., Rockosi, C. M., Yanny, B., & Lee, Y. S. 2006, *ApJ*, 636, 804
- Aoki, T. E., Hiromoto, N., Takami, H., & Okamura, S. 1991, *Publ. Astron. Soc. Jpn.*, 43, 755
- Baade, W. 1944, *ApJ*, 100, 137
- Bahcall, J. N., & Casertano, S. 1985, *ApJL*, 293,7
- Bahcall, J. N., & Casertano, S. 1986, *ApJ*, 308, 347
- Bahcall, J. N., & Soneira, R. M. 1980, *ApJS*, 44, 73
- Bahcall, J. N., & Soneira, R. M. 1984, *ApJS*, 55, 67
- Bahcall, J. N., Ratnatunga, K. U., Buser, R., Fenkart, R. P., & Spaenhauer, A. 1985, *ApJ*, 299, 616
- Bartkevicius, A., Bartasiute, S., & Lazauskaite, R. 1997, in *Proc. ESA Symp., Hipparcos '97*, ed. B. Battrock, M. A. C. Perryman & P. L. Bernacca (Noordwijk: ESA), 343
- Becker, W. 1980, *A&A*, 87, 80
- Beers, T. C., & Sommer-Larsen, J. 1995, *ApJS*, 96, 175
- Bensby, T., Feltzing, S., & Lundström, I. 2003, *A&A*, 410, 527
- Bensby, T., Feltzing, S., & Lundström, I. 2004, *A&A*, 415, 155
- Bensby, T., Feltzing, S., Lundström, I., & Ilyin, I. 2005, *A&A*, 433, 185
- Bensby, T., Zenn, A. R., Oey, M. S., & Feltzing, S. 2007, *ApJL*, 663,13
- Bica, E., Bonatto, C., Barbuy, B., & Ortolani, S. 2006, *A&A*, 450, 105
- Binney, J., Gerhard, O., & Spergel, D. 1997, *MNRAS*, 288,365
- Blaauw, A. 1965, in *Galactic Structure*, ed. A. Blaauw & M. Schmidt (Chicago: The University of Chicago Press), 435
- Boehringer, G., & Hensler, G. 1989, *A&A*, 215, 147
- Bond, N. A., Ivezić, Ž., Sesar, B., Jurić, M., et al. 2010, *ApJ*, 716, 1
- Bournaud, F., Elmegreen, B. G., & Martig, M. 2009, *ApJL*, 707, 1
- Brook, C. B., Kawata, D., Gibson, B. K., & Freeman, K. C. 2004, *ApJ*, 612, 894
- Brown, W. R., Beers, T. C., Wilhelm, R., Allende Prieto, C., Geller, M. J., Kenyon, S. J., & Kurtz, M. J. 2008, *AJ*, 135, 564
- Burkert, A., & Yoshii, Y. 1996, *MNRAS*, 282, 1349
- Burkert, A., Truran, J. W., & Hensler, G. 1992, *ApJ*, 391, 651
- Burstein, D. 1979, *ApJ*, 234, 829
- Buser, R., Rong, J., & Karaali, S. 1998, *A&A*, 331, 934
- Buser, R., Rong, J., & Karaali, S. 1999, *A&A*, 348, 98
- Cabrera-Lavers, A., Garzón, F., & Hammersley, P. L. 2005, *A&A*, 433, 173
- Carney, B. W., Latham, D. W., & Laird, J. B. 1989, *AJ*, 97, 423
- Carollo, D., Beers, T. C., Lee, Y. S., et al. 2007, *Nature*, 450, 1020
- Carollo, D., Beers, T. C., Chiba, M., Norris, J. E., Freeman, K. C., Lee, Y. S., Ivezić, Ž., Rockosi, C. M., & Yanny, B. 2010, *ApJ*, 712, 692
- Carretta, E., & Gratton, R. G. 1997, *A&A Suppl.*, 121, 95
- Casetti-Dinescu, D. I., Girard, T. M., Korchagin, V. I., & van Altena, W. F. 2011, *ApJ*, 728, 7
- Chaboyer, B., Sarajedini, A., & Armandroff, T. E. 2000, *AJ*, 120, 3102
- Chen, B. 1997, *AJ*, 113, 311
- Chen, B., Stoughton, C., Smith, J. A., et al. 2001, *ApJ*, 553, 184
- Chiba, M., & Beers, T. C. 2000, *AJ*, 119, 2843
- Chiu, L.-T. G. 1980, *ApJS*, 44, 31
- Cohen, M. 1995, *ApJ*, 444, 874
- Collins, M. L. M., Chapman, S. C., Ibata, R. A., Irwin, M. J., Rich, R. M., Ferguson, A. M. N., Lews, G. F., Tanvir, N., & Koch, A. 2011, *MNRAS*, 413, 1548
- Conti, P. S., & Vacca, W. D. 1990, *AJ*, 100, 431
- Cox, D. P. 1983, *ApJL*, 265, 61
- Croswell, K., Latham, D. W., Carney, B. W., Schuster, W., & Aguilar, L. 1991, *AJ*, 101, 2078
- Da Costa, G. S. 1982, *AJ*, 87, 990
- Dahn, C. C., Liebert, J., Harris, H. C., & Guetter, H. H. 1995, in *Proc. ESO Workshop, The Bottom of the Main Sequence – and Beyond*, ed. C. G. Tinney (Berlin: Springer), 239
- Dawson, P. C. 1986, *ApJ*, 311, 984
- de Grijs, R. 1998, *MNRAS*, 299, 595
- de Grijs, R., & van der Kruit, P. C. 1996, *A&A Suppl.*, 117, 19
- de Grijs, R., Peletier, R. F., & van der Kruit, P. C. 1997, *A&A*, 327, 1997
- del Rio, G., & Fenkart R. 1987, *A&A Suppl.*, 68, 397
- de Vaucouleurs, G., & Pence, W. D. 1978, *AJ*, 83, 1163
- Dierickx, M., Klement, R., Rix, H.-W., & Liu, C. 2010, *ApJL*, 725, 186
- Digby, A. P., Hambly, N. C., Cooke, J. A., Reid, I. N., & Cannon, R. D. 2003, *MNRAS*, 344, 583
- Drimmel, R., & Spergel, D. N. 2001, *ApJ*, 556, 181
- Du, C., Zhou, X., Ma, J., Bing-Chih Chen, A., Yang, Y., Li, J., Wu, H., Jiang, Z., & Chen, J. 2003, *A&A*, 407, 54
- Du, C., Zhou, X., Ma, J., Shi, J., Chen, A. B.-C., Jiang, Z., & Chen, J. 2004, *AJ*, 128, 2265
- Duquenoay, A., & Mayor, M. 1991, *A&A*, 248, 485

- Edvardsson, B., Andersen, J., Gustafsson, B., Lambert, D. L., Nissen, P. E., & Tomkin, J. 1993, *A&A*, 275, 101
- Elvius, T. 1965, in *Galactic Structure*, ed. A. Blaauw & M. Schmidt (Chicago: The University of Chicago Press), 41
- Feltzing, S., Bensby, T., & Lundström, I. *A&A Lett.*, 2003, 297, 1
- Fenkart, R., Topaktas L., Boydag S., & Kandemir, G. 1987, *A&A Suppl.*, 67, 245
- Fischer, D. A., & Marcy, G. W. 1992, *ApJ*, 396, 178
- Freeman, K. C. 1970, *ApJ*, 160, 811
- Fuhrmann, K. 1998, *A&A*, 338, 161
- Fukuoka, T., & Simoda, M. 1976, *Publ. Astron. Soc. Jpn.*, 28, 633
- Gilmore, G., 1984, *MNRAS*, 207, 223
- Gilmore, G., & Reid, N. 1983, *MNRAS*, 202, 1025
- Gilmore, G., & Wyse, R. F. G. 1985, *AJ*, 90, 2015
- Gilmore, G., & Wyse, R. F. G. 1986, *Nature*, 322, 806
- Gilmore, G., Reid, I. N., & Hewett, P. C. 1985, *MNRAS*, 213, 257
- Gilmore, G., Wyse, R. F. G., & Kuijken, K. 1989, *ARA&A*, 27, 555
- Girard, T. M., Korchagin, V. I., Casetti-Dinescu, D. I., van Altena, W. F., López, C. E., & Monet, D. G. 2006, *AJ*, 132, 1768
- Gizis, J. E. 1997, *AJ*, 113, 80
- Gliese, W. 1969, *Catalogue of Nearby Stars*, Edition 1969, Vol. 22 (Heidelberg: Veröfentl. Astron. Rechen-Inst)
- Gould, A. 2003, *ApJ*, 583, 765
- Gould, A., Bahcall, J. N., & Flynn, C. 1996, *ApJ*, 465, 759
- Gratton, R. G., & Ortolani, S. 1989, *A&A*, 211, 41
- Gratton, R. G., Fusi Pecci, F., Carretta, E., Clementini, G., Corsi, C. E., & Lattanzi, M. 1997, *ApJ*, 491, 749
- Gratton, R. G., Bragaglia, A., Carretta, E., Clementini, G., Desidera, S., Grundahl, F., & Lucatello, S. 2003, *A&A*, 408, 529
- Hammersley, P. L., Cohen, M., Garzón, F., Mahoney, T., & López-Corredoira, M. 1999, *MNRAS*, 308, 333
- Harris, W. E. 1976, *AJ*, 81, 1095
- Hartkopf, W. I., & Yoss, K. M. 1982, *AJ*, 87, 1679
- Hartwick, F. D. A. 1970, *ApJ*, 161, 845
- Hartwick, F. D. A. 1976, *ApJ*, 209, 418
- Hawkins, M. R. S. 1984, *MNRAS*, 206, 433
- Helmi, A. & White, S. D. M. 1999, *MNRAS*, 307, 495
- Helmi, A., White, S. D. M., de Zeeuw, P. T., & Zhao, H. 1999, *Nature*, 402, 53
- Hesser, J. E., Harris, W. E., Vandenberg, D. A., Allwright, J. W. B., Shott, P., & Stetson, P. B. 1987, *Publ. Astron. Soc. Pac.*, 99, 739
- Ivezić, Ž., Sesar, B., Jurić, M., et al. 2008, *ApJ*, 684, 287
- Jahreiß, H., & Wielen, R. 1997, in *Proc. ESA symp., HIPPARCOS '97*, ed. B. Battrock, M. A. C. Perryman & P. L. Bernacca (Noordwijk: ESA), 675
- Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, *ApJ*, 673, 864
- Just, A., & Jahreiß, H. 2010, *MNRAS*, 402, 461
- Karaali, S., Bilir, S., & Hamzaoglu, E. 2004, *MNRAS*, 355, 307
- Katz, N. 1992, *ApJ*, 391, 502
- Kron, R. G. 1980, *ApJS*, 43, 305
- Kroupa, P. 1995, *ApJ*, 453, 350
- Kuijken, K., & Gilmore, G. 1989, *MNRAS*, 239, 605
- Larsen, J. A. 1996, PhD thesis, University of Minnesota
- Larsen, J. A., & Humphreys, R. M. 2003, *AJ*, 125, 1958
- Lin, D. N. C., & Pringle, J. E. 1987a, *MNRAS*, 225, 607
- Lin, D. N. C., & Pringle, J. E. 1987b, *ApJL*, 320, 87
- Liu, W. M., & Chaboyer, B. 2000, *ApJ*, 544, 818
- Lopez-Corredoira, M. 2006, *MNRAS*, 369, 1911
- Lynden-Bell, D., & Pringle, J. E. 1974, *MNRAS*, 168, 603
- Majewski, S. R. 1992, *ApJS*, 78, 87
- Majewski, S. R. 1993, *ARA&A*, 31, 575
- McCuskey, S. W. 1966, *Vistas Astron.*, 7, 141
- Meléndez, J., Asplund, M., Alves-Brito, A., et al. 2008, *A&A Lett.*, 484, 21
- Miller, G. E., & Scalo, J. M. 1979, *ApJS*, 41, 513
- Mishenina, T. V., Soubiran, C., Kovtyukh, V. V., & Korotin, S. A. 2004, *A&A*, 418, 551
- Ng, Y. K., Bertelli, G., Chiosi, C., & Bressan, A. 1997, *A&A*, 324, 65
- Norris, J. 1986, *ApJS*, 61, 667
- Norris, J. E., & Green, E. M. 1989, *ApJ*, 337, 272
- O'Connell, D. J. K. 1958, in *Ricerche Astronomiche, Specola Vaticana*, Vol. 5, ed. D. J. K. O'Connell (Amsterdam: North-Holland)
- Ojha, D. K. 2001, *MNRAS*, 322, 426
- Ojha, D. K., Bienaymé, O., Mohan, V., & Robin, A. C. 1999, *A&A*, 351, 945
- Olivier, S. S., Blumenthal, G. R., & Primack, J. R. 1991, *MNRAS*, 252, 102
- Oort, J. 1960, *Bull. Astron. Inst. Neth.*, 15, 45
- Ortiz, R., & Lepine, J. R. D. 1993, *A&A*, 279, 90
- Paust, N. E. Q. Chaboyer, B., & Sarajedini, A. 2007, *AJ*, 133, 278
- Perryman, M. A. C., Lindegren, L., Kovalevsky, J., et al. 1995, *A&A*, 304, 69
- Plaut, L. 1965, in *Galactic Structure*, ed. A. Blaauw & M. Schmidt (Chicago: The University of Chicago Press), 267
- Pohlen, M., Balcells, M., Lütticke, R., & Dettmar, R.-J. 2004, *A&A*, 422, 465
- Porcel, C., Garzon, F., Jimenez-Vicente, J., & Battaner, E. 1998, *A&A*, 330, 136
- Pritchett, C. 1983, *AJ*, 88, 1476

- Prochaska, J. X., Naumov, S. O., Carney, B. W., McWilliam, A., & Wolfe, A. M. 2000, *AJ*, 120, 2513
- Raghavan, D., McAlister, H. A., Henry, T. J., Latham, D. W., Marcy, G. W., Mason, B. D., Gies, D. R., White, R. J., & ten Brummelaar, T. A. 2010, *ApJS*, 190, 1
- Ramírez, I., Allende Prieto, C., & Lambert, D. L. 2007, *A&A*, 465, 271
- Reddy, B. E., Lambert, D. L., & Allende Prieto, C. 2006, *MNRAS*, 367, 1329
- Reid, M. J. 1993, *ARA&A*, 31, 345
- Reid, I. N. 1997, *AJ*, 114, 161
- Reid, I. N., & Gizis, J. E. 1997, *AJ*, 113, 2246
- Reid, I. N., & Majewski, S. R. 1993, *ApJ*, 409, 635
- Reid, I. N., Gizis, J. E., & Hawley, S. L. 2002, *AJ*, 124, 2721
- Riaz, B., Gizis, J. E., & Samaddar, D. 2008, *ApJ*, 672, 1153
- Robin, A. C., & Crézé, M. 1986, *A&A Suppl.*, 64, 53
- Robin, A. C., Haywood, M., Crézé, M., Ojha, D. K., & Bienaymé, O. 1996, *A&A*, 305, 125
- Robin, A. C., Reylé, C., & Crézé, M. 2000, *A&A*, 359, 103
- Rood, R. T., Carretta, E., Paltrinieri, B., Ferraro, F. R., Fusi Pecci, F., Dorman, B., Chieffi, A., Straniero, O., & Buonanno, R. 1999, *ApJ*, 523, 752
- Ruchti, G. R., Fulbright, J. P., Wyse, R. F. G., et al. 2011, *ApJ*, 737, 9
- Ruphy, S., Robin, A. C., Epchtein, N., Copet, E., Bertin, E., Fouque, P., & Guglielmo, F. 1996, *A&A Lett.*, 131, 21
- Ryan, G. R., & Norris, J. E. 1991, *AJ*, 101, 1865
- Saio, H., & Yoshii, Y. 1990, *ApJ*, 363, 40
- Sales, L. V., Helmi, A., Abadi, M. G., et al. 2009, *MNRAS Lett.*, 400, 61
- Sandage, A. 1957, *ApJ*, 125, 422
- Sandage, A. 1987, *AJ*, 93, 610
- Sandage, A., & Fouts, G. 1987, *AJ*, 93, 592
- Sandage, A., Lubin, L. M., & Vandenberg, D. A. 2003, *Publ. Astron. Soc. Pac.*, 115, 1187
- Schmidt, M. 1959, *ApJ*, 129, 243
- Schmidt, M. 1975, *ApJ*, 202, 22
- Schönrich, R., & Binney, J. 2009, *MNRAS*, 399, 1145
- Schönrich, R., Asplund, M., & Casagrande, L. 2011, *MNRAS*, 415, 3807
- Seth, A. C., Dalcanton, J. J., & de Jong, R. S. 2005, *AJ*, 130, 1574
- Shigejima, T., & Tsujimoto, T. 1998, *ApJ*, 507, 135
- Siegel, M. H., Majewski, S. R., Reid, I. N., & Thompson, I. B. 2002, *ApJ*, 578, 151
- Siegel, M. H., Karata, Y., & Reid, I. N. 2009, *MNRAS*, 395, 1569
- Silk, J., & Norman, C. 1981, *ApJ*, 247, 59
- Simoda, M., & Fukuoka, T. 1976, *Publ. Astron. Soc. Jpn.*, 28, 641
- Sommer-Larsen, J., & Christensen, P. R. 1989, *MNRAS*, 239, 441
- Spagna, A., Lattanzi, M. G., Lasker, B. M., McLean, B. J., Massone, G., & Lanteri, L. 1996, *A&A*, 311, 758
- Spagna, A., Lattanzi, M. G., Re Fiorentin, P., & Smart, R. L. 2010, *A&A*, 510, 4
- Stetson, P. B., & Harris, W. E. 1988, *AJ*, 96, 909
- Stobie, R. S., & Ishida, K. 1987, *AJ*, 93, 624
- Toomre, A. 1964, *ApJ*, 139, 117
- Trefzger, Ch. F., Pel, J. W., & Blaauw, A. 1983, in *The Milky Way Galaxy*, ed. H. van Woerden, R. J. Allen, & W. E. Burton (Dordrecht: Reidel), 151
- Trefzger, Ch. F., Pel, J. W., & Gabi, S. 1995, *A&A*, 304, 381
- Tritton, K. P., & Morton, D. C. 1984, *MNRAS*, 209, 429
- Tsujimoto, T., Nomoto, K., Yoshii, Y., Hashimoto, M., Yanagida, S., & Thielemann, F.-K. 1995, *MNRAS*, 277, 945
- Tsujimoto, T., Shigejima, T., & Yoshii, Y. 1999, *ApJL*, 519, 63
- Tyson, J. A., & Jarvis, J. F. 1979, *ApJL*, 230, L153
- Ugoren, A. R. 1963, *AJ*, 68, 475
- Vandenberg, D. A., Swenson, F. J., Rogers, F. J., Iglesias, C. A., & Alexander, D. R. 2000, *ApJ*, 532, 430
- van der Kruit, P. C. 1984, *A&A*, 140, 470
- van der Kruit, P. C. 1988, *A&A*, 192, 117
- van der Kruit, P. C., & Searle, L. 1981, *A&A*, 95, 105
- van der Kruit, P. C., & Searle, L. 1982, *A&A*, 110, 61
- Veltz, L., Bienaymé, O., Freeman, K. C., et al. 2008, *A&A*, 480, 753
- Villalobos, Á., & Helmi, A. 2008, *MNRAS*, 391, 1806
- Vivas, A. K., & Zinn, R. 2006, *AJ*, 132, 714
- von Hippel, T., & Bothun, G. D. 1993, *ApJ*, 407, 115
- Wainscoat, R. J., Freeman, K. C., & Hyland, A. R. 1989, *ApJ*, 337, 163
- Wielen, R. 1974, *Highlights Astron.*, 3, 395
- Wielen, R., Jahreiß, H., & Krüger, R. 1983, in *Proc. IAU Colloquium 76, The Nearby Stars and the Stellar Luminosity Function*, ed. A. G. Davis Philip & A. R. Ugoren (New York, L. Davis), 163
- Wilson, M. L., Helmi, A., Morrison, H. L., et al. 2011, *MNRAS*, 413, 2235
- Wyse, R. F. G., & Gilmore, G. 1989, *Comments Astrophys.*, 13, 135
- Wyse, R. F. G., Gilmore, G., Norris, J. E., Wilkinson, M. I., Kleyna, J. T., Koch, A., Evans, N. W., & Grebel, E. K. 2006, *ApJL*, 639, 13
- Yamagata, T., & Yoshii, Y. 1992, *AJ*, 103, 117
- Yamagata, T., & Yoshii, Y. 1994, in *Proc. IAU Symp.* 161, ed. H. T. MacGillivray, E. B. Thomson, B.

- M. Lasker, I. N. Reid, D. F. Malin, R. M. West, & H. Lorenz (Dordrecht: Kluwer), 420
- Yanny, B., Newberg, H. J., Grebel, E. K., et al. 2003, *ApJ*, 588, 824
- Yoachim, P., & Dalcanton, J. J. 2006, *AJ*, 131, 226
- Yoshii, Y. 1982, *Publ. Astron. Soc. Jpn*, 34, 365
- Yoshii, Y. 1984, *AJ*, 89, 1190
- Yoshii, Y., & Sommer-Larsen, J. 1989, *MNRAS*, 236, 779
- Yoshii, Y., Ishida, K., & Stobie, R. S., 1987, *AJ*, 93, 323
- Yoshii, Y., Tsujimoto, T., & Nomoto, K. 1996, *ApJ*, 462, 266
- Yoss, K. M., Neese, C. L., & Hartkopf, W. I. 1987, *AJ*, 94, 1600
- Young, P. J. 1976, *AJ*, 81, 807
- Zinn, R., & West, M. J. 1984, *ApJS*, 55, 45

9 The Infrared Galaxy

Ed Churchwell¹ · Robert A. Benjamin²

¹Department of Astronomy, University of Wisconsin, Madison, WI, USA

²Department of Physics, University of Wisconsin - Whitewater, Whitewater, WI, USA

1	<i>The Infrared Era of Galactic Astronomy</i>	449
1.1	A Survey of Surveys	450
2	<i>Stellar Content and Structure of the Galaxy</i>	454
2.1	An Overview of Infrared Galactic Stellar Surveys	454
2.2	The Stellar Disk	458
2.2.1	Distance to the Galactic Center	458
2.2.2	Scalelength(s)	459
2.2.3	Spiral Structure	460
2.2.4	The Stellar Warp, Flare, and Cutoff	462
2.3	The Galactic Bar(s)	464
2.3.1	The Long Bar	464
2.3.2	Inner Bar?	465
2.3.3	The Inner Hole (and Ring?)	466
3	<i>Interstellar Dust</i>	466
3.1	Spatial Distribution of Extinction	467
3.2	Wavelength Dependence of Extinction (Mid-infrared)	467
3.3	PAHs Emission	468
3.4	Stochastic and Thermal Dust Emission	472
4	<i>Star Formation</i>	473
4.1	Infrared Dark Clouds	474
4.2	Extended Green Objects	476
4.3	Massive Young Stellar Objects and the Galactic Star Formation Rate	477
4.3.1	PAH Bubbles and Triggered Star Formation	479
4.4	Massive Star Formation Regions: A Case Study	482
5	<i>Evolved Stars</i>	485
5.1	Variable Stars	485
5.2	Asymptotic Giant Stars	487
5.3	Planetary Nebulae	487

5.4	Luminous Blue Variables and Wolf-Rayet Stars	488
5.5	Supernova Remnants	489
6	<i>Limitations and Lessons Learned</i>	489
	<i>Acknowledgments</i>	491
	<i>References</i>	491

Abstract: As infrared surveys have reached optical-quality angular resolution, they have revealed new information on the stellar, interstellar, and star-formation components of the Galaxy. The distance to the Galactic center appears to be known to within 5%: $R_o = 8.0 \pm 0.4$ kpc. Measurements of the stellar scalelength of the disk, $R_d = 2-4$ kpc, continue to show a large range; the origin of this scatter needs to be understood. The exponential disk does not continue into the center of the Galaxy, with an inner radius of $R_h \sim 3$ kpc. Claims exist for a truncation, or change in scalelength, in the outer disk, but are not yet confirmed. The stellar disk is warped, with a similar nonsymmetric azimuthal dependence as the HI disk, but a lower amplitude and uncertain radial extent. There is extensive evidence for two non-axisymmetric structures in the inner galaxy: the Galactic bar (or triaxial bulge) and the Long Bar, which differ in angle by $\sim 20^\circ$. The existence of an inner (nuclear) bar seems likely, but studies have not converged on its parameters. There is no compelling evidence for a ring in stellar mass, but a case can be made for a star-forming ring.

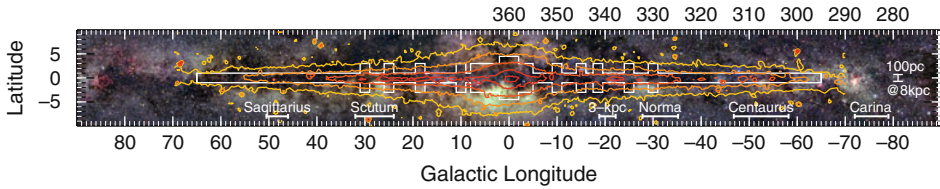
These surveys have also revealed the projected Galactic distribution of interstellar dust and shed light on the extinction and radiation properties of dust in different Galactic environments. We review the spatial distribution and wavelength dependence of extinction, the distribution and magnitude of PAH emission and stochastic and thermal dust emission, and different stages of massive star formation, e.g., infrared dust clouds (IRDCs), IR bubbles around HII regions, extended green objects (EGOs), and massive star formation regions, including evidence for triggered star formation. We briefly discuss what infrared observations tell us about evolved stars, including variable stars, asymptotic giant stars (AGB), planetary nebulae (PNe), Luminous Blue Variables (LBVs) and Wolf-Rayet stars, and supernova remnants. We end with a comparison of the limitations of optical, infrared, and radio surveys of the Galaxy that should be borne in mind.

Keywords: Galaxy:stellar content, Galaxy:structure, Infrared:general, Infrared:ISM, Infrared:stars, ISM: general, Stars:general, Survey

1 The Infrared Era of Galactic Astronomy

Among the most strikingly beautiful features of an optical panorama of the Milky Way (► [Fig. 9-1](#)) are the prominent dust lanes running along the Galactic plane. But in terms of making progress in understanding the structure of the Milky Way, particularly the stellar structure, dust has been the enemy. Because of the dust and the vast spread in distance of objects in a given direction, less is known about the large-scale structure and kinematics of our Galaxy, the Milky Way, than many nearby spiral galaxies. Determination of the extent and morphology of the Galaxy presents much the same problem faced by a hiker lost in a foggy forest. To determine the extent and shape of the forest, somehow the hiker has to find a way to see all the way to the edge of the forest and estimate distances along the line of sight.

Until the mid-1980s, our knowledge of the extent and structure of the Milky Way was almost exclusively from large-scale radio surveys of HI and CO gas tracers. However, certain fundamental limitations frustrated one of the ultimate goals of these surveys: the creation of a reliable map of the distribution of interstellar gas in the Milky Way. The recent discovery of the *Far Three-Kiloparsec Arm* (Dame and Thaddeus 2008) provides a striking example of these limitations. This structure is comparable in kinematics, atomic and molecular mass to the *Near Three-Kiloparsec Arm* (van Woerden et al. 1957). Its existence was predicted by



■ Fig. 9-1

Optical panorama of the inner Milky Way (Mellinger 2009). The patchy optical emission can be compared with the relatively smooth mid-infrared light observations of COBE/DIRBE $4.9\ \mu\text{m}$, shown in contours of logarithmic intensity (MJy/sr). The area covered by the original GLIMPSE survey, with $|b| < 1^\circ$ except for certain longitude ranges, is shown in white. At a distance of 8 kpc, 1° corresponds to 140 pc. The “historical” longitude ranges for spiral arm tangencies are taken from Englmaier and Gerhard (1999). The main bar (or triaxial bulge) of the Galaxy extends from $l \sim 12^\circ$ to about -8° . The Long Bar extends from $l = 30^\circ$ to about -15°

Oort (1977) and more recent models of gas flow in a barred potential, c.f. Merrifield (2004). Yet it eluded detection for more than half a century! Why? Ultimately, the limitations in the angular resolution of even the most recent surveys, as well as the difficulties of converting longitude-velocity diagrams into longitude-distance diagrams, obscured this structure for many years.

It should also be noted that these pioneering surveys of HI and CO map only one component of the Galaxy: the interstellar gas. Observations of extragalactic systems show that the distribution of interstellar gas, of star formation, and of the stellar mass of a single spiral galaxy can be strikingly different (Block and Wainscoat 1991); their interrelationship contains information on how galaxies work as star formation factories.

In the last three decades, infrared surveys of the Galaxy have begun to shed new light on the global distribution, and differences in the distributions, of these three components of the Galaxy. In this review, we divide the infrared regime into the near-infrared (NIR) bands ($1\text{--}2.5\ \mu\text{m}$), the mid-infrared (MIR) bands ($2.5\text{--}50\ \mu\text{m}$), and the far-infrared (FIR) bands ($50\text{--}500\ \mu\text{m}$). The near-infrared bands contain information about the distribution of stars throughout the galaxy as well as the physical nature and spatial distribution of dust extinction. The far-infrared bands provide constraints on the distribution and physical state of dust emission throughout the Galaxy. The mid-infrared bands have the optimum combination of low extinction and low diffuse emission needed to study the stellar content of the Galaxy, but are also ideal for studying the emission of polycyclic aromatic hydrocarbons (PAHs) and thermal emission from very small dust grains.

1.1 A Survey of Surveys

Since the 1980s, our knowledge of the Galaxy’s structure and content has vastly expanded as a result of the opening of the infrared frontier. Much of this progress has come as a result of numerous surveys. ▶ [Table 9-1](#) contains basic information on these surveys. ▶ [Figure 9-2](#) compares their point source sensitivities. The conversion from flux units (Janskys= Jy) to magnitudes for different near- and mid-infrared surveys is given in ▶ [Table 9-2](#).

Table 9-1
Summary of infrared surveys^a

Survey	Wavebands (μm)	Resolution ($''$)	Coverage	Sensitivity (mJy)	Website
DENIS	0.97, 1.22, 2.16	1-3	$\delta = +2$ to -88°	0.2, 0.8, 2.8	cdsweb.u-strasbg.fr/denis.html
2MASS	1.22, 1.65, 2.16	2	all-sky	0.4, 0.5, 0.6	www.ipac.caltech.edu/2mass
UKIDSS-GPS ^b	1.22, 1.65, 2.16	0.5	$l = -2$ to 107° , 142 to 230° ^c	0.016, 0.023, 0.017	www.ukidss.org
GLIMPSE	3.6, 4.5, 5.8, 8.0	≤ 2	$ \leq 65^\circ$, $ b \lesssim 1^\circ$ ^d	0.2, 0.2, 0.4, 0.4	www.astro.wisc.edu/glimpse
GLIMPSE360	3.6, 4.5	≤ 2	$l = 65^\circ - 255^\circ$, $ b \lesssim 2^\circ$	0.012, 0.018	www.astro.wisc.edu/glimpse
WISE	3.4, 4.6, 12, 22	6, 6, 6, 12	all-sky	0.08, 0.1, 1, 6	wise.ssl.berkeley.edu
MSX	4.1, 8.3, 12, 14, 21	18.3	$l = 0 - 360^\circ$, $ b \leq 5^\circ$	10000, 100, 1100, 900, 200	www.ipac.caltech.edu/ipac/msx
MIPSGAL	24, 70	6, 18	$ = 0 - 65^\circ$, $ b \lesssim 1^\circ$	2, 75	mipsgal.ipac.caltech.edu
ISOGAL	7, 15	6	$ \leq 60^\circ$, $ b \leq 1^\circ$ ^e	15, 10	www-isogal.iap.fr/
IRAS	12, 24, 60, 100	25-100	all-sky	350, 650, 850, 3000	irsa.ipac.caltech.edu/IRASdocs
<i>Akari</i>	8.5, 20, 62.5, 80, 155, 175	5-44	all-sky	20-100	www.ir.isas.ac.jp
<i>Herschel</i> /HI-GAL	70, 170, 250, 350, 500	5, 13, 18, 25, 36	$ = 0 - 60^\circ$, $ b \leq 1^\circ$	18, 27, 13, 18, 15	hi-gal.ifsi-roma.inaf.it/higal
COBE/DIRBE ^f	1.25-240	0.7 $^\circ$	all-sky	0.01-1.0 MJy sr^{-1}	space.gsfc.nasa.gov/astro/cobe

^aSee text for appropriate references for these surveys

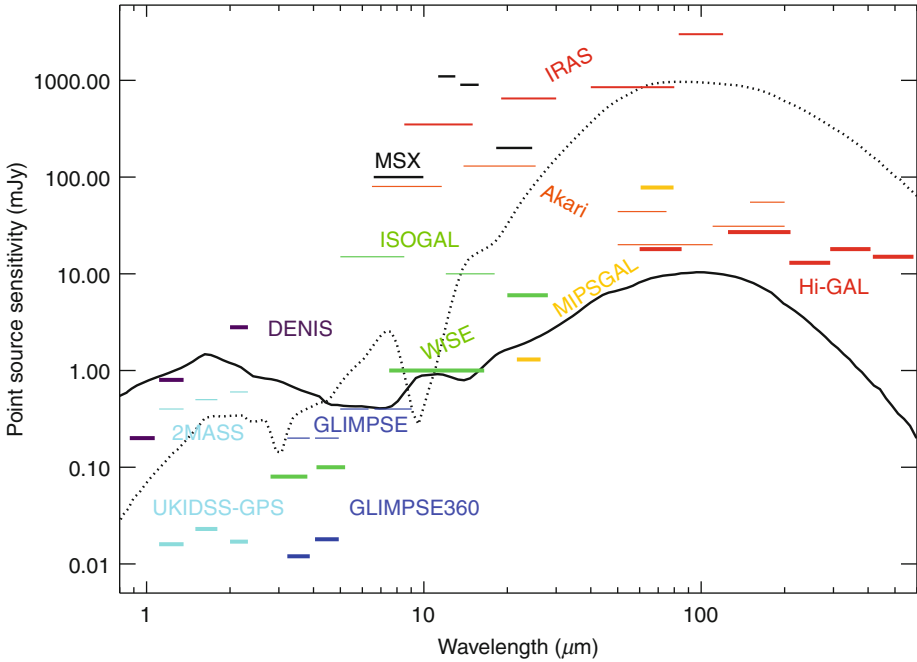
^bMuch of the remainder of the Galactic Plane will be covered with similar depth and resolution in the five-band near-infrared survey VVV (Minniti et al. 2010)

^c $l = -2$ to 15° has thickness $|b| < 2^\circ$, otherwise the thickness is $|b| < 5^\circ$. The longitude range $l = 142^\circ - 230^\circ$ is also covered

^dGLIMPSE also has vertical extensions up to $|b| = 4^\circ$.5 for selected longitudes. GLIMPSE style coverage was used for the *Spitzer* Vela-Carina survey from $l = 295^\circ - 255^\circ$

^eSurvey contained only selected fields in this region, totaling 16 square degrees

^fDIRBE photometric bands are 1.25, 2.2, 3.5, 4.9, 12, 25, 60, 100, 140, and 240 μm . We report the diffuse flux sensitivity rather than point source sensitivity due to the large beam size



■ Fig. 9-2

A comparison of point source sensitivities and wavelength bands of past and current ground and space-based infrared surveys. More details on these surveys are given in [Table 9-1](#). The curves show model spectra of Whitney et al. (2004) for a $1 L_{\odot}$ T Tauri star at a distance of 0.7 kpc (solid) and a deeply embedded $1 L_{\odot}$ protostar at a distance of 0.6 kpc (dotted)

Among the most influential surveys were the IRAS, *Infrared Astronomical Satellite*, survey (Beichman et al. 1988; Neugebauer et al. 1984), the ISOGAL survey of the *Infrared Space Observatory* (Kessler et al. 1996; Omont et al. 2003), MSX, the *Midcourse Space Experiment* (Price et al. 2001), the all-sky COBE/DIRBE, *Cosmic Background Explorer/Diffuse Infrared Background Experiment*, survey (Boggess et al. 1992), 2MASS, *Two Micron All Sky Survey* (Skrutskie et al. 2006), DENIS, *Deep Near Infrared Survey of the Southern Sky* (Fouqué et al. 2000), *Spitzer/GLIMPSE, Galactic Legacy Infrared Mid-Plane Survey Extraordinaire*, (Benjamin et al. 2003; Churchwell et al. 2009), the *Spitzer/MIPSGAL Multiband Imaging Photometer Galactic Plane Survey* (Carey et al. 2009), and the AKARI all-sky survey (Ishihara et al. 2010).

Surveys in progress include the near-infrared surveys UKIDSS-GPS, *United Kingdom Infrared Telescope Deep Sky Survey–Galactic Plane Survey* (Lucas et al. 2008) and VVV, *Vista Variables in the Via Lactea* (Minniti et al. 2010), the mid-infrared all-sky survey WISE, *Wide-field Infrared Survey Explorer* (Wright et al. 2010) and outer Galactic plane survey GLIMPSE 360 (Whitney 2009), and the far-infrared Galactic plane survey, *Herschel/Hi-GAL, Herschel Infrared Galactic Plane Survey* (Molinari et al. 2010).

■ Table 9-2

Effective wavelengths, zero point magnitudes, and extinction in the near- and mid-infrared^a

Band	Wavelength(μm)	S_0	A_λ/A_K
2MASS J	1.235	1594.0	2.50 ± 0.15
2MASS H	1.662	1024.0	1.55 ± 0.08
2MASS K _s	2.159	667.0	1.0
WISE [3.4]	3.353	306.681	...
IRAC [3.6]	3.550	280.9	0.56 ± 0.06
IRAC [4.5]	4.493	179.7	0.43 ± 0.08
WISE [4.6]	4.603	170.663	...
IRAC [5.8]	5.731	115.0	0.43 ± 0.10
IRAC [8.0]	7.872	64.13	0.43 ± 0.10
MSX A	8.276	58.5	...
WISE [12]	11.561	29.0448	...
MSX C	12.126	26.5	...
MSX D	14.649	18.3	...
MSX E	21.336	8.8	...
WISE [22]	22.088	8.2839	...
MIPS [24]	23.68	7.14	...
MIPS [70]	71.42	0.775	...
MIPS [160]	155.9	0.159	...

^a2MASS, IRAC, and MIPS calibration, effective wavelengths, and zero magnitude fluxes are discussed in Rieke et al. (1995) and references therein. MSX values are from Cohen et al. (2001); WISE values are from Wright et al. (2010). Typical extinctions from Indebetouw et al. (2005); see Sect. 3 information on variation with Galactic direction. For reference, $A_{[4.5]} = 0.43A_K \cong 0.05A_V$

Starting with 2MASS (near-infrared) and GLIMPSE (mid-infrared), these surveys have reached optical-quality angular resolutions and sensitivities, making it possible to detect individual stars, as well as better resolve diffuse emission and stellar clusters. This advance in resolution and sensitivity has made it possible to begin a study of the large-scale distribution of stars in the Galaxy, as opposed to the integrated light from multiple stars and diffuse dust emission. The improved angular resolution also allows for a great improvement in our understanding of massive star formation, allowing for the separation of the stars, embedded YSOs, and the complex structure of diffuse emission from PAHs and dust grains.

In this review, we discuss some of the major hallmarks of the Galaxy at infrared wavelengths. In Sect. 2, we collect and compare infrared-based results on the overall stellar structure of the Galaxy, including results from star counts and the use of standard candles. Section 3 covers the spatial distribution, extinction, and emission of interstellar dust and PAHs. In Sect. 4, we outline how infrared studies have improved our understanding of the stages and distribution of massive star formation throughout the Galaxy, leading to new methods of estimating the global star formation of the Milky Way Galaxy. In Sect. 5, we summarize some of the results of infrared investigations of evolved stars and the return of their gas and dust to the interstellar medium. We close our review in Sect. 6 with some final thoughts about the strengths and limitations of infrared investigations of the Galaxy.

The topic covered here is sufficiently vast that this review will inevitably be incomplete. Some discoveries, possibly important, will have been overlooked. Some readers may also find this review somewhat tilted to the results of the large-scale surveys, particularly in the mid-infrared (where the authors have had the most experience), as opposed to smaller area photometric and spectroscopic investigations. Finally, this review is principally focused on the Galactic thin disk and bar(s). This review will not cover infrared emission at high latitudes, e.g., the infrared cirrus or the stellar thick disk, and will only briefly touch on infrared investigations of the Galactic center region.

2 Stellar Content and Structure of the Galaxy

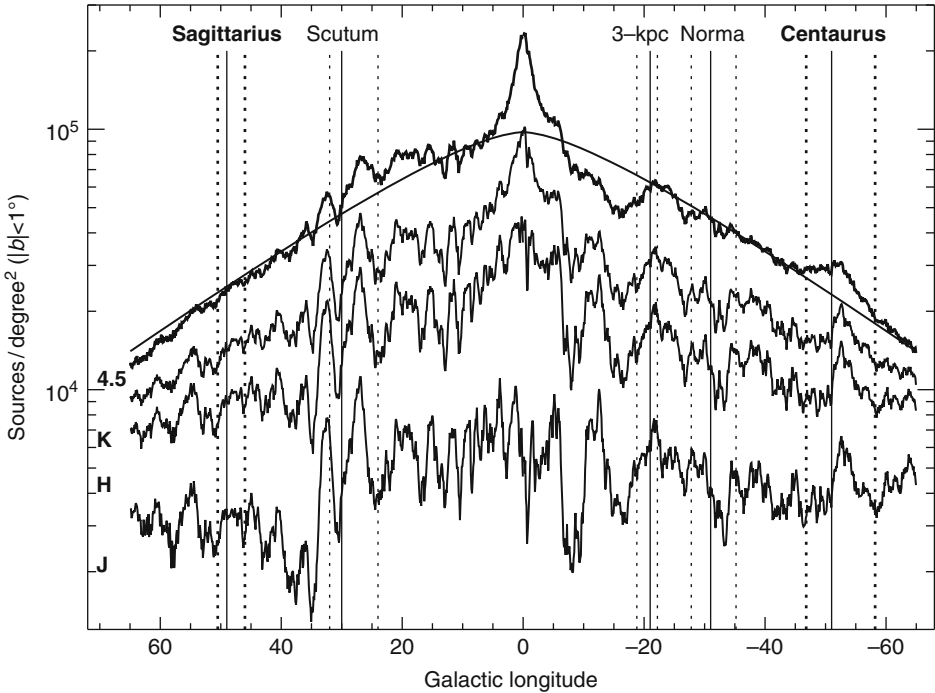
In the midplane of the Galaxy, recent infrared surveys have been detecting tens of millions of objects never before catalogued. This dramatic increase is due to three factors: greater sensitivity, improved angular resolution, and the decreased extinction at longer wavelengths (🔗 Fig. 9-3). Depending on these limitations, as well as sky coverage, different observational programs are sensitive to different “components” of the Galaxy, i.e., bulge, disk, bar, etc. Understanding the observational limitations of these different programs, and the models used to interpret them, is vital to make sense of the disparate measurements of Galactic parameters. The section below discusses some general considerations in sorting through the zoo of available surveys, and then a review of some of the recent work in characterizing the Galactic structure.

2.1 An Overview of Infrared Galactic Stellar Surveys

Optical surveys of the Galaxy, c.f. SEGUE, *Sloan Extension for Galactic Understanding and Exploration* (Yanny et al. 2009), typically detect sources to much fainter magnitudes than current infrared surveys. However, because of the fundamental limit imposed by extinction, optical studies principally constrain the structure of the stellar halo, satellites, and the thick disk. The stellar structure of the thin disk can only be probed in the solar neighborhood, $D < 2$ kpc (Jurić et al. 2008), and optical studies of the bulge or bar are limited to a few low extinction windows a few degrees off the midplane. Surprisingly, where deep optical studies cross the midplane, they can detect many nearby dwarf stars that are *not* seen in infrared surveys. Since infrared surveys have much lower extinction, these faint nearby dwarf stars can be lost against the brighter glare of all the more distant giant stars made detectable by lowered extinction. The effective depth (faint magnitude limit) of infrared surveys can be determined by this *confusion limit*, not just the survey *sensitivity limit*. The confusion limit depends on the angular resolution, extinction, and source density, and is a particular concern in the inner galactic plane, bar, and stellar clusters.

Current infrared surveys of the Galactic midplane are principally surveys of different classes of red giants. This can be seen in 🔗 Fig. 9-4, which shows a TRILEGAL, *TRIdimensional model of thE GALaxy*,¹ simulation (Girardi et al. 2005) in the inner galaxy. One sees a decrease in the luminosity function at an absolute magnitude of $M_K \sim -1$, the break between giant and

¹Girardi et al. (2005) note this word also means “very nice” in southern Brazil. And you thought the acronym GLIMPSE was contrived!

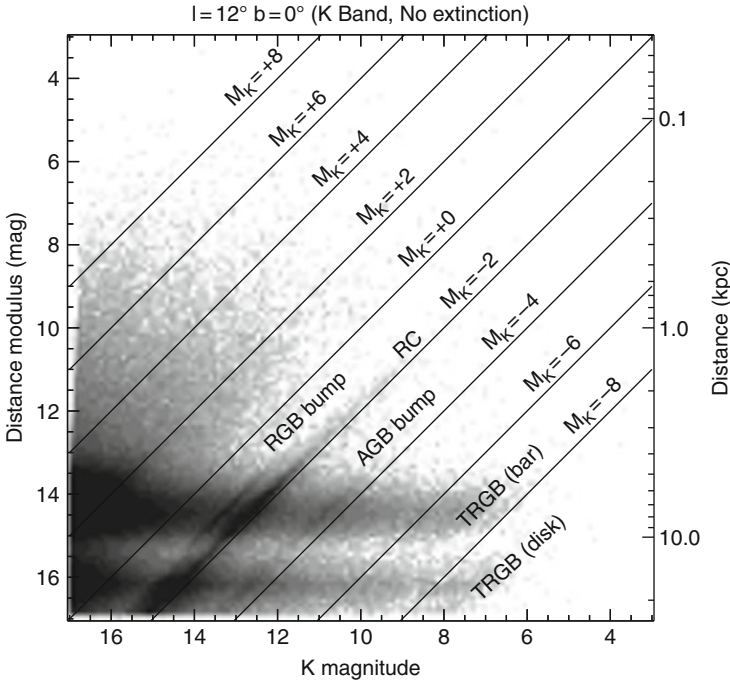


■ Fig. 9-3

Sources per square degree in the magnitude range $m = 6 - 12$ averaged over the latitude strip $|b| < 1^\circ$ binned with a resolution of $\Delta l = 6$ arcminutes. The near infrared data for the J ($1.2 \mu\text{m}$), H ($1.6 \mu\text{m}$), and K ($2.2 \mu\text{m}$) bands are from the 2MASS point source catalog; the mid-infrared [4.5] band data comes from the GLIMPSE point source catalog. Note that star counts increase with increasing wavelength, and the local dips, due to clouds of absorption, become less pronounced. The fit to the mid-infrared star-count data (for $|l| = 30 - 65^\circ$ only) is a first-order modified Bessel function of the second kind, $N = N_0 (l/l_0) K_1 (l/l_0)$, expected for an edge-on exponential disk, and yields $l_0 = 24 \pm 4^\circ$. Things to note include (1) the relative symmetry between positive and negative longitudes, (2) the enhancement in source counts interior to $|l| = 30^\circ$ due to the bar(s), (3) the presence of a “hole” (compared to the expectation of a filled exponential disk) from $l = -8^\circ$ to $\sim -20^\circ$, and (4) a broad excess centered at $l = -52^\circ.6$ in the expected Centaurus spiral arm tangency direction

dwarf stars. The confusion/sensitivity limits of $m \sim 14$ mag for recent near- and mid-infrared surveys (2MASS, GLIMPSE) mean that most of the sources detected in these surveys are giants. Current deeper surveys (UKIDSS-GPS, GLIMPSE360) should also be detecting large numbers of dwarfs in regions that are not confusion-limited.

There are two ways to convert point source catalogs into constraints on Galactic parameters: star-count models and standard candles. Surveys without sufficient angular resolution to resolve stars, e.g., COBE/DIRBE, must necessarily use the first method, convolved with the detector beam-size. The chief advantage of star-count models is that one uses all of the sources, providing excellent statistics. The chief disadvantage, discussed at length in Mihalas and Binney (1981),



■ Fig. 9-4

TRILEGAL prediction for the number of sources as a function of distance modulus and predicted K band magnitude for 1 deg², zero-extinction area centered on $(l, b) = (12^\circ, 0^\circ)$ using the default values of Girardi et al. (2005). The horizontal band at distance modulus, $\mu = 14.5$ mag, is the contribution of the Galactic bulge (Vanhollebeke et al. 2009), while the disk contribution, which depends on the combination of space density and size of the volume element, is maximum at $\mu \sim 16$ mag. Characteristic features of the predicted giant luminosity function are noted. Lines of constant absolute magnitude are oriented diagonally. A predicted source histogram can be obtained by summing vertically along this diagram

is that extraction of Galactic parameters requires specifying (1) a model for both the stellar density as a function of position in the Galaxy and (2) the luminosity function of sources. Since the luminosity function is broad, sources at a given apparent magnitude come from a range of distances. This degrades one's ability to converge on a unique density model for the Galaxy. In addition, the resulting best-fit parameters provide a useful shorthand for summarizing results, but may be misleading in the case of unanticipated structures or degeneracies between model parameters.

There are characteristic features in the giant luminosity function (► Fig. 9-4) that can be used as standard candles to map stellar density along a line of sight. Chief among these features are the *tip of the red giant branch* (TRGB, $M_K \sim -6.85$), *red clump* (RC) stars with a local absolute magnitude of $M_K = -1.54 \pm 0.04$ (Groenewegen 2008), *red giant branch bump* stars (RGB bump, $M_K \sim -1.0$), and *asymptotic giant branch bump* stars (AGB bump, $M_K \sim -3.3$). These features have been detected in globular clusters, old open clusters, and Local Group galaxies. It is thought that the red clump star luminosity is (relatively) insensitive to metallicity or population age, with a maximum spread of about 0.4 magnitudes (Girardi and Salaris 2001). This will affect *absolute*

distance estimates. *Relative* distances when mapping galactic structures should be more secure unless there are large metallicity or population gradients. Other features are probably more dependent on metallicity and age effects. This can be seen in [▶ Fig. 9-4](#) where the TRGB for disk stars is nearly two magnitudes brighter than the TRGB for older, metal-poor bulge stars.

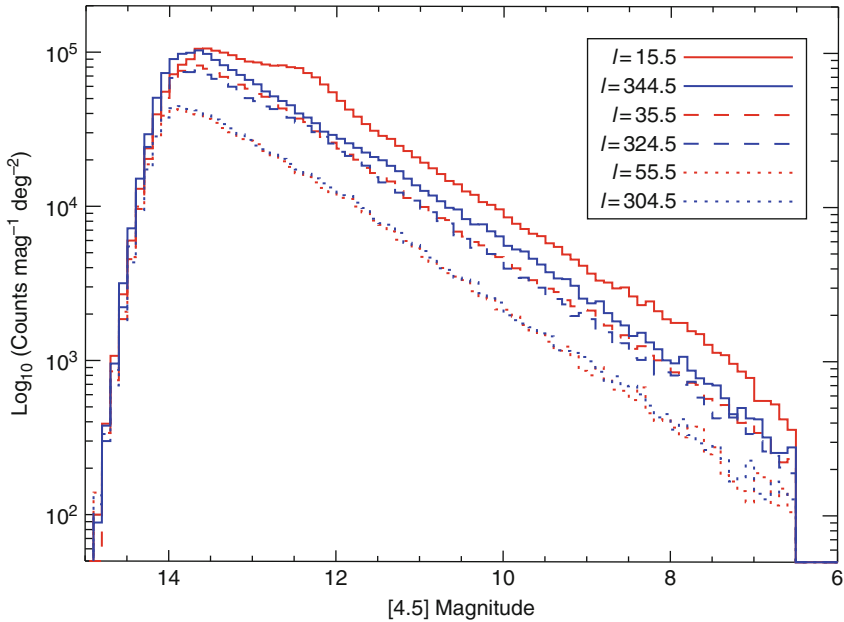
Other, less common, infrared bright sources have also been used to map the inner Galaxy (principally the bars), including carbon stars (Cole and Weinberg 2002), OH/IR stars (Sevenster 1999), AGB star (Weinberg 1992), and Mira variables (Groenewegen and Blommaert 2005). A useful table describing different classes of stars can be found in Chapter 3 of Binney and Merrifield (1998). Since these sources are comparatively rare, the statistics are poorer. However, they provide important constraints on population age and metallicity that might be washed out in samples of ordinary giants. In addition, since they have a much tighter range of magnitudes, they more tightly constrain the density function of the component of the Galaxy being mapped. However, it is very important to remember that sources which trace young populations are essentially mapping the *star-forming structure* of the Galaxy, *not* the mass density.

Given the large number of potential classes of sources that can be used, one approach is to develop models that match all the features of an infrared color-magnitude diagram *simultaneously* (van Loon et al. 2003). This approach has been used with great success for Local Group dwarf galaxies. In the Galaxy, the additional free parameters needed to characterize stellar density and extinction vs position, plus the sheer volume of data to be analyzed, make this approach quite challenging.

Of the four features in the giant luminosity function discussed above, red clump stars have been put to the most use in mapping Galactic structure. This is because they have the highest space density, the tightest luminosity function, and have been absolutely calibrated using a sample with *Hipparcos* parallaxes (Alves 2000; Groenewegen 2008). Identifying red clump stars in the field is more challenging than in clusters or Local Group galaxies because of the effects of extinction and distance spreads along the line of sight. If one wants to obtain a “pure” sample of red clump giants at a given apparent magnitude, one needs color information to separate them out from brighter and fainter red giants (and dwarfs) at the same apparent magnitude.

With mid-infrared data alone, this separation is not possible. Mid-infrared wavelengths sample the Rayleigh-Jeans tail of the spectra of ordinary dwarfs and giants. As a result, the mid-infrared colors of non-dusty dwarfs and giants is near zero. However, the near-infrared colors of red clump giants is slightly bluer, $(J - K_s) = 0.70 \pm 0.05$ (Alves 2000; Grocholski and Sarajedini 2002), than most red giants, and much redder than dwarfs. They can therefore be color selected provided one can make adequate extinction corrections. The number of red clump stars per bin of apparent magnitude can be converted to mass density as a function of distance, assuming the number of red clump stars is proportional to the total mass.

There is one situation in which the distance to red clump stars can be obtained *without* color selection. Whenever there are relative overdensities in the Galaxy, a histogram of sources as a function of magnitude will have a “bump” at a magnitude related to the distance of the overdensity. [▶ Figure 9-5](#) shows an example of mid-infrared histograms showing a bump at $m_{[4.5]} \sim 12.5$ due to red clump stars in the galactic “Long Bar” (Benjamin et al. 2005). If the apparent magnitude of the bump shifts smoothly with longitude, the bump is probably due to a standard candle tracing a region of stellar overdensity in the Galaxy. Red clump stars are the most likely class of source to produce such a bump; near-infrared color magnitudes can be used to confirm this. [▶ Figure 9-6](#) shows the slope of point source histograms as a function of longitude and magnitude for GLIMPSE 4.5 μm data. Several slope changes associated with different Galactic structures are noted and discussed below.



■ Fig. 9-5

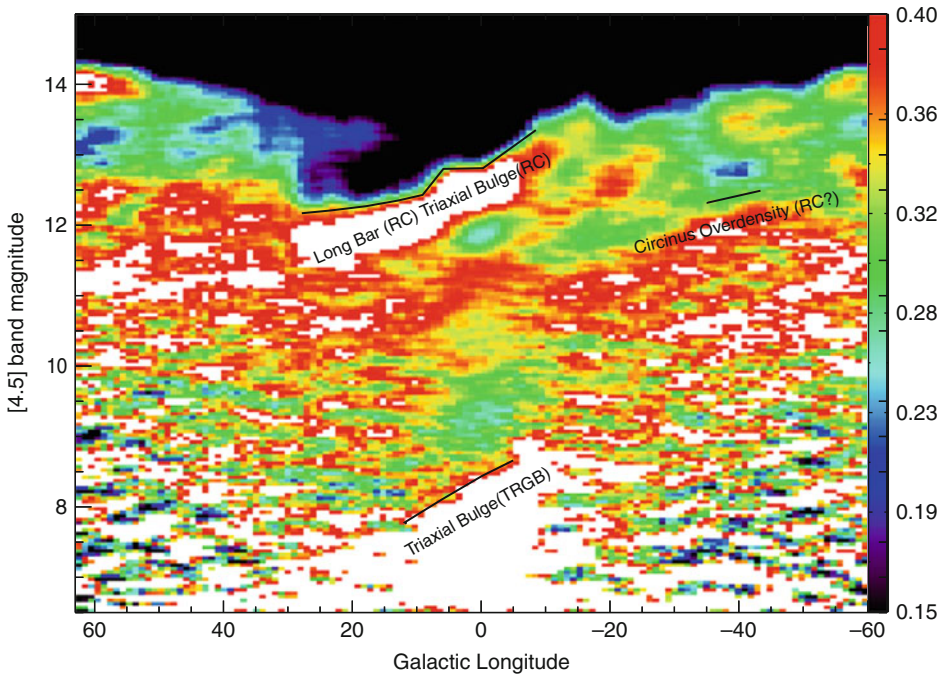
Number of sources from the GLIMPSE Point Source Catalog as a function of magnitude for three pairs of directions. The outer galaxy ($l = 55.5^\circ/l = 304.5^\circ$) and middle galaxy ($l = 35.5^\circ/l = 324.5^\circ$) curves have approximately the same amplitude and slopes. The inner galaxy ($l = 15.5^\circ/l = 324.5^\circ$) shows a significant north/south asymmetry; the northern direction also shows a bump in source counts at a magnitude of $m \sim 12.2$. The number of sources has been averaged over a 1° (longitude) $\times 1^\circ.8$ (latitude) region. Figure from Benjamin et al. (2005)

With the rapid improvement in sensitivity and resolution, researchers are still identifying new classes of sources and developing new analysis techniques to estimate different Galactic parameters. The diversity of approaches and results can be bewildering. Which results are right? Which are suspect? This review will not make those choices. In general, the results that should carry the most weight should be those that are (1) based on all-sky (or all-plane) surveys, (2) have resolved stellar populations, (3) have the lowest intrinsic extinction, and (4) use well-calibrated standard candles.

2.2 The Stellar Disk

2.2.1 Distance to the Galactic Center

The distance to the Galactic center, R_0 , is a key parameter for galactic structure. Although the official IAU value remains 8.5 kpc (Kerr and Lynden-Bell 1986), the bulk of measurements since then have favored somewhat lower values (Reid 1993). The distance to red clump stars in the bulge indicate an even smaller distance of $7.52^{+0.35(\text{sys})}_{\pm 0.1(\text{stat})}$ kpc (Nishiyama et al. 2006). The two most reliable methods to get the distance to the Galactic center are infrared observations to



■ Fig. 9-6

Slope of the Spitzer/GLIMPSE source histograms, $\log N$ vs magnitude, shown in Fig. 9-5 as a function of magnitude and longitude. Color bar at the right shows the value of the slope. Density enhancements of red clump giants along a line of sight produce changes in the slope, with a magnitude that changes with longitude. Slope changes due to red clump stars in the Long Bar, the triaxial bulge, and the Circinus overdensity (possibly the Scutum-Centaurus spiral arm) are noted. The slope change in source counts in the inner galaxy around eighth magnitude is due to the tip of the red giant branch in the triaxial bulge

determine the orbital parameters of the S2 star around Sgr A*, $R_0 = 8.4 \pm 0.4$ kpc (Ghez et al. 2008) or $R_0 = 8.33 \pm 0.35$ kpc (Gillessen et al. 2009), and measurements of the radio parallax of the nearby Sgr B2, $R_0 = 7.9^{+0.8}_{-0.7}$ kpc (Reid et al. 2009). As observations continue for these two objects, the statistical errors will decrease. In the following sections, we have not attempted to correct reported measurements to put them on the same R_0 scale.

2.2.2 Scalelength(s)

A definitive determination of the scalelength of the Galactic stellar thin disk has yet to be made, and we are not aware of a recent critical review of the various measurements. Older studies, summarized by Robin et al. (1992b), yielded values in the range $R_d = 3.5\text{--}4.5$ kpc. Subsequent investigations, summarized by Sackett (1997), have usually yielded smaller values, $R_d = 2.5\text{--}3.0$ kpc. Since this summary, six additional optical studies have measured values ranging from 2.25 to 4.0 kpc, c.f. Jurić et al. (2008) and references therein.

Infrared studies should, in principle, give a more reliable measure of the thin disk scalelength as they can probe further through the disk in the midplane. In the inner galaxy ($|l| < 90^\circ$), one key test of the reliability of the results is that they should be symmetric on either side of Galactic center outside the longitude range affected by the bar. In the outer galaxy, measurement of the scalelength is complicated by the presence of flaring (increase in scaleheight with radius) and warping. Results based on full-sky COBE/DRIBE low angular resolution observations of infrared light range from 2.4 to 2.6 kpc (Freudenreich 1998) and 2.3 kpc (Drimmel and Spergel 2001).

Near-infrared data from DENIS or 2MASS have also been used to constrain the scalelength, although most of these attempts have excluded the inner Galaxy. This includes analysis of DENIS data for $l = 217^\circ$ and $l = 239^\circ$, $R_d = 2.3 \pm 0.1$ kpc (Ruphy et al. 1996) and analysis of 2MASS data from $l = 90\text{--}270^\circ$, $R_d = 2$ kpc (Reyl   et al. 2009). L  pez-Corredoira et al. (2002) used both star-count models and red clump giants for selected 2MASS fields in the longitude range $45^\circ < l < 315^\circ$ to estimate the scalelength, finding agreement between the two methods. Because the scaleheight was found to change with radius, the scalelength of the midplane density, $R_{d,0} = 2.0$ kpc, and the scalelength of the surface stellar mass density, $R_{d,tot} = 2.4$ kpc, differ. Finally, because of the reduced extinction in the mid-infrared, GLIMPSE star-count data, combined with the luminosity function of Wainscoat et al. (1992), constrained the scalelength in the *inner* Galaxy, $|l| = 30\text{--}65^\circ$, yielding $R_{d,0} = 3.9 \pm 0.6$ kpc (Benjamin et al. 2005).

Making sense of all of these results would be a valuable project. As Binney and Tremaine (2008) note, given other constraints on the Galactic potential, the difference between a scalelength of 2 kpc and 3.2 kpc is the difference between a stellar mass dominated gravitational potential, and one dominated by dark matter. Some of the difficulties in comparing the results of different authors include the following: (1) observations of other disk galaxies show the radial scalelength to be wavelength dependent; (2) radial variation in the scaleheight of the thin disk and the presence of the thick disk, with a separate scalelength and scaleheight, produces potential degeneracies; (3) other galaxies show evidence of two separate scalelengths for the outer disk, depending on Hubble type (Erwin et al. 2008); and (4) surveys using photometric distances to sources need to account for binary stars (Juri   et al. 2008).

2.2.3 Spiral Structure

Spiral structure in disk galaxies is seen most prominently as a pattern of gas density and star formation. This pattern can be well organized in grand-design spirals or patchy in flocculent spirals. But disk galaxies also show spiral structure in the stellar mass, traced by infrared light. The morphology and amplitudes of the mass spiral and the star-formation spiral can be notably different (Schweizer 1976; Zwicky 1955), although they bear some similarities. The difference in arm morphology as a function of tracer is discussed at length in Chapter 6 of Binney and Tremaine (2008), who suggest the names *mass arm*, *potential arm*, *gas arm*, and *bright-star arm* to distinguish the organization of different tracers. (In this review, we refer to *star formation arm* as opposed to *bright-star arm*.) In general, the mass distribution, which is traced by the near- and mid-infrared light, is smoother and less structured than the distribution of gas or star formation, which is traced by blue light. It is even possible for the number of spiral arms seen in mass to differ from the number of arms seen in star formation (Block et al. 2004; Block and Wainscoat 1991). The idea that a two-armed mass spiral could drive more than two star

formation arms was first suggested by Shu et al. (1973); recent models intended to be applicable to the Milky Way are presented in Martos et al. (2004).

Determining the spiral structure of the Galaxy using tracers of gas and star formation has been a challenging problem (Liszt 1985) due to uncertainties in kinematic distances to gas clouds and HII regions and the uncertainties in photometric distances to the bright stars in the arms. One would expect that the spiral arm tangency directions would be more secure, but unambiguous identification of even these directions has been problematic (► Fig. 9-7). The common picture of four primary arms, *Norma*, *Sagittarius-Carina*, *Scutum-Crux*,² and *Perseus*, characterized by HII regions (Georgelin and Georgelin 1976), has come to dominate the literature, with other structures relegated to secondary status. These secondary features include the *Near Three Kiloparsec* arm (van Woerden et al. 1957), the *Far Three Kiloparsec Arm* (Dame and Thaddeus 2008), the *Orion Spur*³ (Morgan et al. 1953), the *Outer Arm*⁴ (Westervhout 1957), and the *Distant Arm* (McClure-Griffiths et al. 2004). The distinction between primary and secondary features is based on the number of bright HII regions thought to lie in each structure; a reexamination of this with modern data would be very informative. There is, unfortunately, no recent critical review of spiral structure in the Milky Way; we feel the current picture should still be considered provisional. ► Figure 9-8 shows an artist's schematic of the Galaxy which contains most of these features at approximately the correct longitudes and estimated distances.

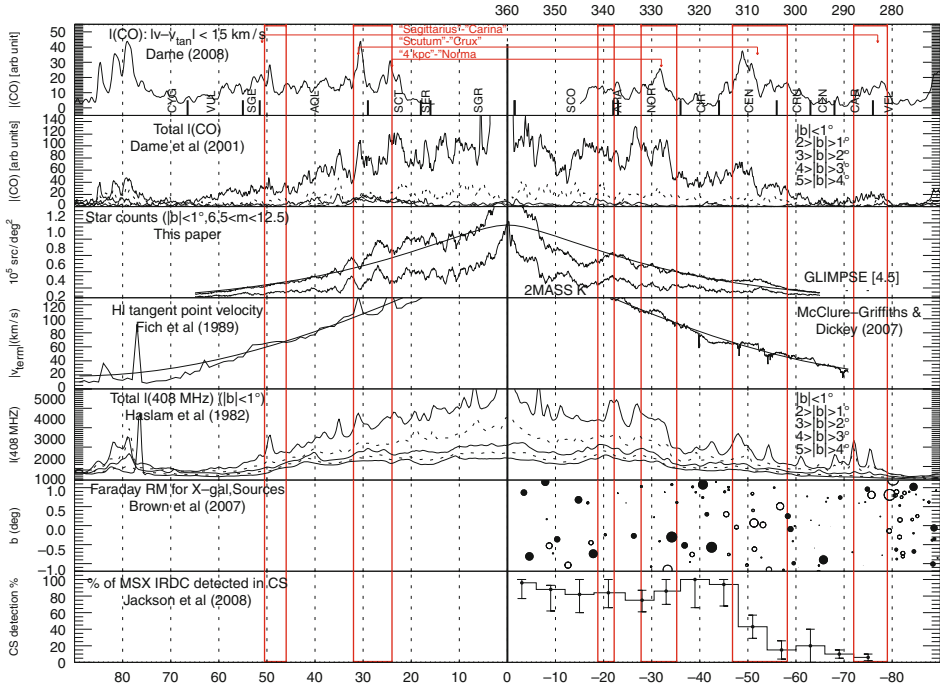
The spiral structure of the mass distribution of the Galaxy may be simpler. Based on near-infrared studies, mass arms are enhanced by 0.2–0.6 with respect to the azimuthal average (Rix and Zaritsky 1995) and typical cross sections (width at a given circular radius) of 20–40° (Seigar and James 1998). These widths are narrower than would be expected for a simple sinusoidal density variation, which would predict a FWHM of 60° for a two-armed spiral. This discrepancy may indicate either (1) a deficiency in the single-mode model of spiral structure or (2) indicate that the near-infrared light distribution may be affected by star formation (Rhoads 1998) as well as mass distribution.

An analysis of the K band light distribution of COBE/DIRBE suggests that Galactic spiral structure stellar mass is qualitatively different from the gas and star formation picture described above (Drimmel 2000; Drimmel and Spergel 2001). The principal evidence for this is the detection of an enhancement in near-infrared light in the direction of the Centaurus tangency, but no corresponding enhancement in the $l \sim 50^\circ$ direction of the Sagittarius Arm tangency. This observational result is confirmed with mid-infrared star counts (Benjamin et al. 2005), which show an ~30% enhancement in star counts (of all magnitudes) centered at $l = 307^\circ$ with a full-width at half-max of 4°. Models by Drimmel and Spergel (2001) using COBE/DIRBE near-infrared light to constrain the stellar mass distribution and the far-infrared light to constrain the dust distribution found that the data were consistent with two principal mass arms (Perseus and Scutum-Crux) and four arms in gas/star formation. The newest infrared surveys may allow for the direct *mapping* of the mass arms using red clump giants or other standard candles. ► Figure 9-6 shows evidence of an enhancement in mid-infrared star counts at $l = 316\text{--}326^\circ$ and $m_{[4.5]} \sim 12.3$; this longitude range is in a gap in the CO distribution (Dame et al. 2001). If this is due to red clump giants, the feature would be (approximately) consistent with the expected distance to the Scutum-Crux arm.

²Also called *Scutum-Centaurus*

³Also called the *Orion Arm*, *Local Arm*, or *Cygnus Arm*

⁴Also called the *Cygnus Arm*, or in a speculative leap, the *Norma-Cygnus* arm

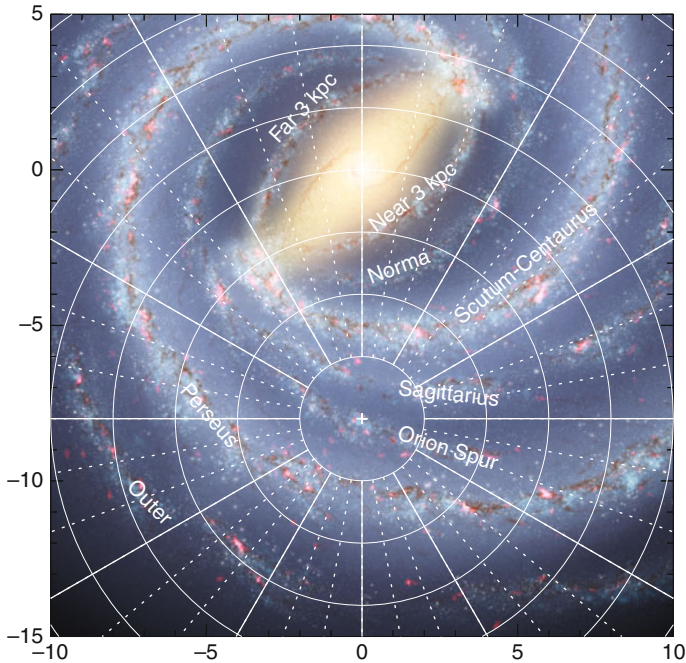


■ Fig. 9-7

A comparison of different tracers that might be expected to show spiral arm tangencies. From the bottom up, this includes high-density molecular clouds detected in CS (Jackson et al. 2008), Faraday rotation of extragalactic sources (Brown et al. 2007), radio synchrotron emission with different latitude cuts (Haslam et al. 1982), variations in the tangent point velocity measured in HI (Fich et al. 1989; McClure-Griffiths and Dickey 2007), mid- and near-infrared star counts (Benjamin et al. 2005), total CO intensity with different latitude cuts (Dame et al. 2001), and the integrated CO intensity within 15 km/s of the tangent point velocity. The longitude ranges for the expected tangency directions are shown and paired together for each spiral arm. Except for the top plot, there does not appear to be a compelling correlation between the data and expected tangency directions. However, the Centaurus tangency direction does appear to contain an enhancement in CO, stars, a reversal in Faraday rotation measure, and marks a significant drop in the number of infrared dark clouds that contain CS

2.2.4 The Stellar Warp, Flare, and Cutoff

An interesting and unsolved question is how to measure (or even define) the edge of galactic disks. In the Galaxy, it is not yet clear which component of the galactic disk extends to the greatest distance, the interstellar gas or the stars. What is clear, however, is that both components of the Galaxy flare (increasing vertical scaleheight with radius) and warp (changing midplane with azimuth). Evidence of the stellar warp from COBE/DIRBE data (Freudenreich et al. 1994; Freudenreich 1998; Drimmel and Spergel 2001) showed that the parameters characterizing the stellar warp were similar in phase to the gaseous (neutral hydrogen) warp (Burton et al. 1992), but smaller in amplitude. More recent attempts based on near- and mid-infrared star counts



■ Fig. 9-8

An artist's conception of the Milky Way Galaxy based on the data summarized here. Details on the construction of this image can be found in Churchwell et al. (2009). These features show most of the main structures of the Galaxy with approximately correct longitudes and distances. Circles, centered on the Sun, are located every two kiloparsecs, and Galactic longitude is marked every 10° , with $l = 0^\circ$ oriented upward and $l = 90^\circ$ to the left. Much of the structure beyond the distance of Galactic center is extremely speculative and assumes bi-symmetry

(López-Corredoira et al. 2002; Reylé et al. 2009; Vig et al. 2005) or using red clump stars as standard candles (Momany et al. 2006) have confirmed this general behavior, although there are pronounced asymmetries in the amplitude of the warp with azimuth.

The overall agreement of the gaseous and stellar warp show that the warp is due principally to gravitational as opposed to gas dynamical effects. However, the reason behind the difference in the overall amplitudes, as well as the asymmetries, is as yet unclear. Clearly, the different radial distributions of gas and stars will play a role. Extragalactic studies, e.g., Erwin et al. (2008), have found exponential outer disk profiles that are steeper or shallower than the inner disk profiles, depending on the Hubble type, although this may be a result of changes in color rather than surface density (Bakos et al. 2008). The transition point occurs around four inner-disk scalelengths, and historically had been thought to mark a disk truncation (van der Kruit and Searle 1981). Evidence for such a truncation at $R = 14$ kpc was reported by Robin et al. (1992a), but some of the references above detect disk stars to even greater radius.


2.3 The Galactic Bar(s)

Although gas kinematics in the inner galaxy have long been pointed to as evidence that the Galaxy has a central non-axisymmetric structure, it was not until the advent of infrared astronomy that it became possible to characterize the stellar structure of the inner Galaxy (Kent et al. 1991). Early results (Blitz and Spergel 1991) provided solid evidence of a stellar Galactic bar. COBE/DIRBE maps were then analyzed by several groups to derive constraints on the Galactic bar. The results are reviewed by Gerhard (2002) and Merrifield (2004). Most of these works tended to find a bar half-length of $R_b = 3.1\text{--}3.5$ kpc oriented in the Galactic plane at $\phi_b \sim 25^\circ$ in the first quadrant with respect to the Galactic center direction. Published values for this angle range from 10° to 40° ; a partial table of values from the literature is given by Vanhollebeke et al. (2009). The bar axis ratios is 10:4:3 (length:width:height), with ranges from 10:7:4 to 10:3:3. The total stellar mass of this structure is $M_b \sim 10^{10} M_\odot$, with estimates ranging from $0.5 - 2 \times 10^{10} M_\odot$. The variations in these estimates are due to differences in the density fitting function used, the difference in luminosity functions, decisions on which other parameters (like the disk component) were held fixed or allowed to float, and incomplete sampling of the sky. The characterization of this long suspected structure, sometimes also referred to as the triaxial bulge, was one of the early highlights of Galactic infrared astronomy.

However, these studies had two disadvantages. First, the relatively poor resolution required assumptions about the unresolved stellar populations to interpret the light distribution (Dwek et al. 1995). It also made bar/bulge decomposition problematic, and meant that thin structures were poorly resolved. Second, although extinction is much lower than optical in the near infrared, it is not negligible, particularly in the midplane. As higher angular resolution infrared surveys started to yield a wealth of information about the distribution of individual types of stars across the Galaxy, the picture has become more complex. Many of these studies traced the same structure as described above. But as near- and mid-infrared surveys allowed the characterization of the stellar populations directly in the midplane, an additional structure became apparent.

2.3.1 The Long Bar

This structure, now referred to as the Long Bar, was first identified as distinct from the bulge by Hammersley et al. (2000). The dawning realization of its existence is detailed in López-Corredoira et al. (2007). Initially characterized in selected low-extinction windows in the near-infrared, the mid-infrared GLIMPSE survey showed that it could be traced continuously with red clump giants out to a longitude of $l \sim 29^\circ.5$ (Benjamin et al. 2005). As compared to the triaxial bulge, this structure is not only longer, $R_{lb} = 3.9$ kpc, but thinner, both in vertical scaleheight, ~ 200 pc, and depth along the line of sight, $\sim 1,170$ pc. The estimated mass of this component, assuming it is symmetric and independent of the central triaxial bulge, is $M_{lb} = 0.6 \times 10^9 M_\odot$ (López-Corredoira et al. 2007). Independent estimates of these parameters would be valuable.

One mystery of this structure is the distinct difference in angle for the Long Bar and the (other) Bar. This difference is not due to different measurement methods. Using red clump giant stars and UKIDSS-GPS data (also see  Fig. 9-6), the same technique shows that the measured bar angle changes from $\phi_{lb} = 42 \pm 2^\circ$ in the longitude range $l = 12 - 30^\circ$ to $\phi_b = 24 \pm 2^\circ$ for $l = 5 - 12^\circ$. At the transition point, the star counts and number of red clump giants jump

sharply; as a result, it is not entirely clear whether the Long Bar continues to exist as a discrete structure inside $l = 12^\circ$. Although other galaxies show evidence of multiple misaligned bars, it is not known how many of these cases are similar to that inferred for the Galaxy.

The significance of the Long Bar is in the realization that a non-axisymmetric mass distribution of the Galaxy extends further from the Galactic center than generally realized. In particular, the near end of the Long Bar coincides with the Scutum spiral arm tangency, a fact first noted by Weinberg (1992). In CO longitude-velocity plots, this direction also marks the first-quadrant tangency point of the Molecular Ring (Dame et al. 2001; Jackson et al. 2006). This intersection appears to be the site of prodigious amounts of star formation, including three large-scale height, $b = 10 - 20^\circ$, superbubbles, c.f. Pidopryhora et al. (2007). Follow-up spectroscopy of the brightest member of GLIMPSE-identified clusters (Mercer et al. 2005) have revealed several clusters of red supergiants, c.f. Clark et al. (2009), one of which is estimated to have a total mass of $20,000 M_\odot$ (Alexander et al. 2009). There is also tentative evidence for star formation at the far end of the Long Bar at $l \sim 345^\circ$ (López-Corredoira et al. 2001; Sevenster 1999). This deserves further scrutiny.

The discovery of the Long Bar also points out a fundamental weakness of parametric modeling of Galactic structure. As numerical simulations of bars show, bars are not monolithic structures, but complex families of stellar orbits whose properties evolve over time. By forcing Galactic models to fit a single parameterized model (based on observations of bars in distant galaxies), we destroy information on the structural complexity of bars. Numerical simulations that start with bars as vertically thin structures eventually produce a thick inner distribution, similar to what is observed in the Galaxy (Athanasoula 2007; Debattista and Shen 2007). It is not clear whether these models can also reproduce the apparent mismatch in angles.


One way to separate the different bar components as well as test for the presence, and star formation history, of a classical or pseudobulge (Kormendy and Kennicutt 2004) is to obtain information on the kinematics and metallicity of individual stars. We do not review the status of these efforts due to lack of space and expertise on the part of the reviewers. A good starting point would be the papers by Zoccali (2010) and Babusiaux et al. (2010). We note, however, that much of this work has been in optical wavelengths, which may give misleading or incomplete results due to the limitations of extinction.

2.3.2 Inner Bar?

Observations of other galaxies show up to *three* nested bars (Erwin et al. 2008), so apparently there is room for one more bar in the Galaxy. Evidence for an inner (sometimes called nuclear) bar, or at least an inner non-axisymmetric structure, has been suggested by the noncircular orbits of gas in the inner few degrees of the Galaxy (Binney et al. 1991), extinction-corrected 2MASS star counts (Alard 2001), and a combination of both of the above (Rodríguez-Fernández and Combes 2008). This final work suggests a bar mass of $M_{nb} = 0.2 \times 10^{10} M_\odot$ and a bar angle, based on the kinematic modeling, of $\phi_{nb} = 60 - 75^\circ$ relative to the Sun-Galactic center direction. This angle is marginally consistent with their independent estimate based on star counts. Other relevant studies that constrain this structure are Launhardt et al. (2002), van Loon et al. (2003), and Sawada et al. (2004). In addition, red clump giant mapping by Nishiyama et al. (2005) shows a flattening of the magnitude vs longitude track of red clump giants for $l = -5^\circ$ to $+5^\circ$. Based on

the evidence so far, it seems that an inner, $|l| < 2^\circ$, mass asymmetry is likely, but its parameters have yet to be firmly established.

2.3.3 The Inner Hole (and Ring?)

The Milky Way is a barred spiral galaxy, but is it a *ringed* barred spiral? Many of the same references that characterized the stellar distribution of the disk and bar (and  Fig. 9-3) require the presence of a central “hole.” Strictly speaking, this is a deficit in stellar density compared to an extrapolation of the exponential disk into the center of the Galaxy. The hole radius is typically 2.7–3.3 kpc with an ellipticity of 0.8–0.9, e.g., Freudenreich (1998). Many barred galaxies are found to have central holes (Ohta et al. 1990); such galaxies are also referred to as Freeman Type II disks (Freeman 1970). A depression in the stellar density inside the radius of the bar is a natural consequence of bar formation, but so far as we know, there has been no detailed comparison between the current dynamical models and the Galactic data constraining the properties of the bar *and* “hole.”

Whether the Galaxy contains a stellar ring, i.e., an *overdensity* of stars surrounding the bar, is considerably more uncertain. There is no compelling evidence for a stellar ring in the *mass* distribution, although obtaining constraints on the ring is complicated by the presence of the bar in the same longitude range. However, there are two structures that could arguably be identified as an interstellar, star-forming ring. The first is the “Molecular Ring” (Burton et al. 1975; Jackson et al. 2006; Scoville and Solomon 1975). First discovered in the first quadrant of the Galaxy, it is clear that a significant fraction of the molecular gas and star formation of the Galaxy is located in an annulus of radius 3–5 kpc. However, it is very difficult to distinguish between a spiral and ring. Several authors have expressed doubts about whether this feature should be interpreted as a ring, c.f. Binney and Merrifield (1998) or Jackson et al. (2008).

Another structure that might plausibly be visible as a star-forming ring to an outside observer is the Near/Far Three-Kiloparsec Arm (Dame and Thaddeus 2008). Simulations of gas flow in a gravitational potential consistent with the Galactic bar (Bissantz et al. 2003; Englmaier and Gerhard 1999; Fux 1999) predict that these two structures should form an oval around the bar. Although early searches of this structure showed no evidence for star formation (Lockman 1980), recent surveys of class II methanol masers show approximately 20 sources in each arm (Green et al. 2009), indicating the presence of massive star formation. An excess of OH/IR stars (Sevenster 1999) and enhanced near-infrared star counts of bright stars, $m_K < 9$ (López-Corredoira et al. 2001), have been noted at $l = 338^\circ$. Both authors noted the coincidence of this direction with the tangency direction of the Near Three-Kiloparsec Arm and suggested that this might be part of a ring, even before the discovery of the Far 3 kpc Arm! A comprehensive model combining the constraints on the mass *and* star formation distribution of this structure, in comparison with the variety of rings seen in other barred spirals (Buta and Combes 1996), would be a worthy endeavor.

3 Interstellar Dust

The distribution and physical state of dust grains with Galactic environment is a vast topic. The most complete review can be obtained in the review by Draine (2003), the monograph by Whittet (2003), or the dust-related chapters in the textbooks of Tielens (2005) or Kwok (2007).

Here we limit our review to recent results in the near- and mid-infrared. The coming deluge of information on longer wavelength dust emission from the *Herschel* and *Planck* missions means that this will be a rapidly advancing field in the next several years.

3.1 Spatial Distribution of Extinction



The distribution of dust throughout the Galaxy, like the distribution of gas, is clumpy. Gas clouds have been traditionally mapped using kinematic distances. But given their relatively smooth distribution, stars could also serve as good tracers of the dust distribution, provided that the extinction to the stars can be reliably estimated. A succinct review of the history and current status of this approach can be found in Robin (2009). The goal is to obtain a three-dimensional map of dust clouds with the highest possible angular and distance resolution, since the extinction is known to be very patchy, and some dust clouds may be quite distant (and therefore appear small).

Drimmel et al. (2003) proposed a three-dimensional extinction model based on their parametric model of the stellar and interstellar distribution constrained by COBE/DIRBE studies. Marshall et al. (2006) assumed a smooth stellar distribution taken from the Besançon model (Robin et al. 2003) and then adjusted extinction vs distance to match the observed 2MASS color magnitude diagrams, obtaining an angular resolution of 15 arcminutes and ~ 100 pc in distance. López-Corredoira et al. (2002) used the distribution of color magnitude selected red clump giants to map extinction along the line of sight, a technique that was also adopted by Durant and van Kerkwijk (2006) and Stead and Hoare (2010). Each of these approaches has different assumptions regarding the background distribution of stars, and therefore can be expected to have different systematic biases. However, because of the patchy nature of the distribution, and different angular and distance resolutions, it is not straightforward to compare the different model predictions. Both Marshall et al. (2009) and Stead and Hoare (2010) compare their extinction distances to the kinematic distances, finding the two distances are correlated. However, Marshall et al. (2009) find that extinction distances in the fourth quadrant were systematically ~ 2 kpc more distant than kinematic distances.

In principle, extinction distances can provide a much needed check on kinematic distances, particularly for gas in the spiral arms and bars, where one might suspect that circular rotation is a poor assumption. Marshall et al. (2006) shows evidence of spiral structure in the distribution of dust clouds. Within three kiloparsecs of Galactic center, Marshall et al. (2008) compare this technique with the results obtained using noncircular kinematic models of gas flow in a barred potential, showing that both techniques give the same distance to clouds that they interpret as dust lanes in the Galactic bar.

3.2 Wavelength Dependence of Extinction (Mid-infrared)

A good overview of wavelength dependence of extinction may be found in Draine (2003). Here we will only update the results obtained in the mid-infrared since that review. Lutz et al. (1996) used ISO spectral line observations to derive the diffuse interstellar extinction law using infrared hydrogen recombination lines toward the Galactic center. They found that A_λ/A_V is essentially constant from ~ 4 to $8 \mu\text{m}$. Jiang et al. (2006, 2003) used ISOGAL and DENIS data toward an 0.1 deg^2 area centered at $l = 18^\circ.6$, $b = 0^\circ.35$ and found a similar flattening of extinction as Lutz et al. (1996). Indebetouw et al. (2005) confirmed these early results using GLIMPSE stellar photometry of red clump giants toward three other lines of sight.

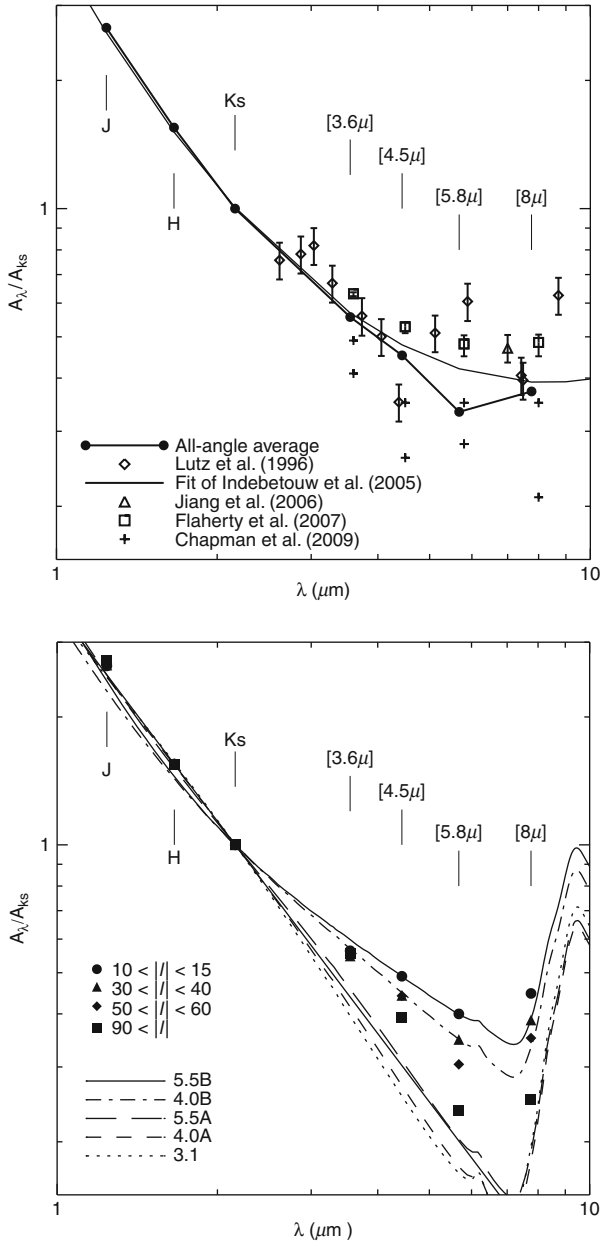
Recently, both Gao et al. (2009) and Zasowski et al. (2009) have conducted a systematic study of the MIR extinction law from 1.2 to 8 μm over a large range of Galactic longitudes, and compare their results to the single-direction extinction measurements made by several groups, including those works mentioned above. The study of Zasowski et al. (2009) used about 150° of contiguous longitude at the Galactic midplane using data from 2MASS and Spitzer surveys (GLIMPSE, Vela-Carina, and Argo). They isolated a sample of red clump giants in J vs (J-K_s) color-magnitude diagrams, and determined the color excess ratios (CER) of this sample in (H- λ) vs (H-K_s) for the *Spitzer*/IRAC and 2MASS bands. The extinction law, A_λ/A_{K_s} , in their formulation depends only on A_H/A_{K_s} and CER_λ . They used $A_H/A_{K_s} = 1.55$ for all longitudes, noting this quantity is likely to be a function of Galactocentric radius. The average change in extinction with longitude from 5.8 to 8.0 μm is shown in  Fig. 9-9. They interpret the increase of $A_{[8.0]}/A_{[5.8]}$ with longitude as evidence for a decrease in mean dust grain size toward the outer Galaxy, noting the Galactic abundance gradient may play a role as well. The grain size dependence is motivated by the theoretical dust models of Weingartner and Draine (2001), which are reproduced in  Fig. 9-9 along with the observed/derived values as a function of longitude.

In contrast, using the same data and a slightly different selection criterion for red clump giants Gao et al. (2009) found an approximately constant value of A_λ/A_{K_s} with galactic longitude, with possible deviations from this constant value in the direction of spiral arm tangencies. When they used red giants, rather than red clump giants, to measure the extinction, A_λ/A_{K_s} was systematically higher, by about ~ 0.05 . This is presumably because red giants span a larger range of distances. Although both works find that A_λ/A_{K_s} decreases with increasing wavelength, the values found by Gao et al. (2009) are systematically higher than those found by Zasowski et al. (2009). It is not clear why one group finds a longitude dependence for the extinction law while the other group does not.

3.3 PAHs Emission

One of the most notable features of mid-infrared images of the Galactic plane is the highly structured diffuse emission. In the GLIMPSE/MIPSGAL surveys, the structure of the diffuse emission changes notably with wavelength, as the relative contributions from unresolved stellar emission, scattered starlight, atomic and molecular emission, and thermal dust emission changes. In the [3.6] and [4.5] *Spitzer*/IRAC bands (where the number is the central wavelength), much of the diffuse emission is due to unresolved point sources (particularly in the inner Galaxy), scattered starlight, and thermal dust emission (particularly in the vicinity of hot stars and regions of massive star formation). The [3.6] band also contains a polycyclic aromatic hydrocarbon (PAH) emission feature at 3.3 μm .

The diffuse emission in the [5.8] and [8.0] bands primarily traces the distribution of PAHs with minor contributions from thermal dust emission near hot stars and stochastic emission from very small dust grains (VSGs) transiently heated by UV photons. Examination of the GLIMPSE/MIPSGAL mosaic images shows that the PAH emission in these bands is widely distributed in the Galactic disk. This emission is brightest at the midplane and declines rapidly with latitude. There are also bright spikes centered on massive star formation regions, many of which are off the plane. Superimposed on the overall variation, the PAH diffuse emission is also characterized by numerous bubbles (Churchwell et al. 2007, 2006) and filamentary structures that may be the result of interstellar turbulence (Heitsch et al. 2007).



■ Fig. 9-9

[Top] Mid-IR reddening as a function of wavelength along different lines of sight through the disk of the Galaxy from different groups noted in the figure. This figure illustrates both the differences between the different groups and the general similarity between them, in particular, the flattening from ~ 4.5 to $8.0 \mu\text{m}$. [Bottom] The observed reddening with wavelength as a function of Galactic longitude for $10^\circ < |l| < 90^\circ$. Several models from Weingartner and Draine (2001) are also plotted showing a change in dust properties that might explain the observed change in the reddening law with longitude. Both figures are from Zasowski et al. (2009)

Figure 9-10 (top) shows the diffuse emission along the Galactic midplane in a residual, i.e., point-source subtracted, [8.0] band image for different latitude cuts. With the exception of the bright spike at the Galactic center and many massive star formation region spikes, the PAH diffuse emission is almost constant, with an average value of ~ 75 MJy/ster, within $\sim 30^\circ$ of the Galactic center, falling off rapidly for $|l| > 30^\circ$. The [8.0] diffuse emission peaks at $b = 0^\circ$ and is tightly confined to the plane. This distribution implies that the inner Galactic plane is permeated by PAHs, whether the decline in emission with longitude traces a decrease in the PAH abundance or a change in the soft (far-ultraviolet) radiation or both is unclear. A detailed characterization of this diffuse emission has been done by Robitaille et al. (2012). A comparison of the Galactic emission with the results of PAH distribution in other galaxies (Draine et al. 2007) could be very informative.

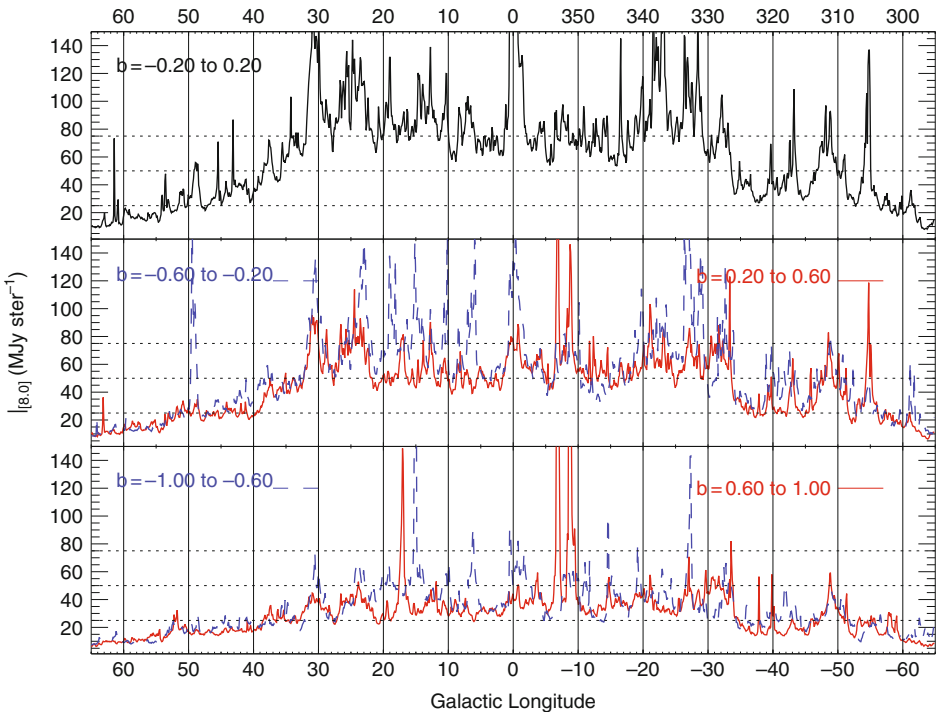
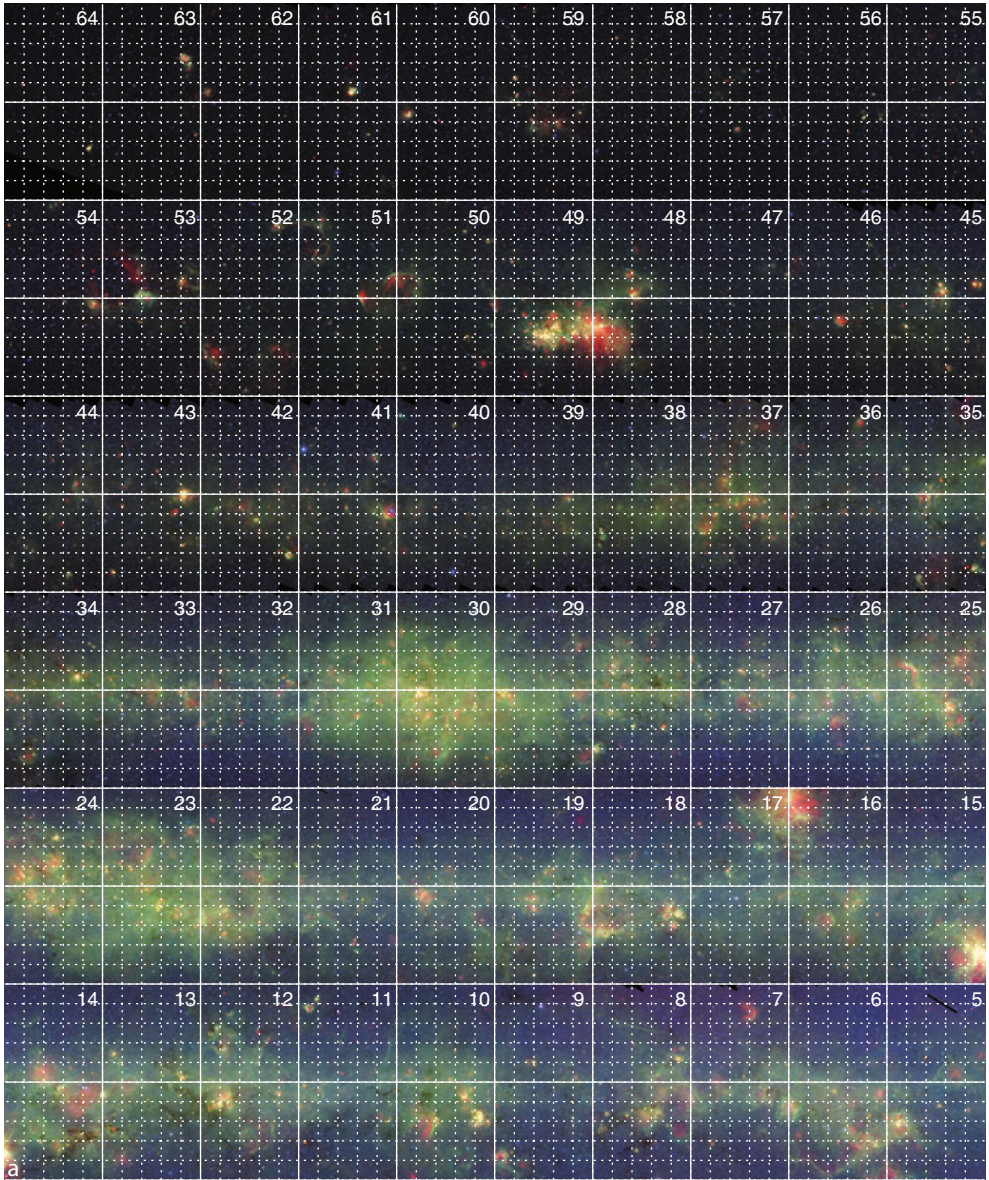


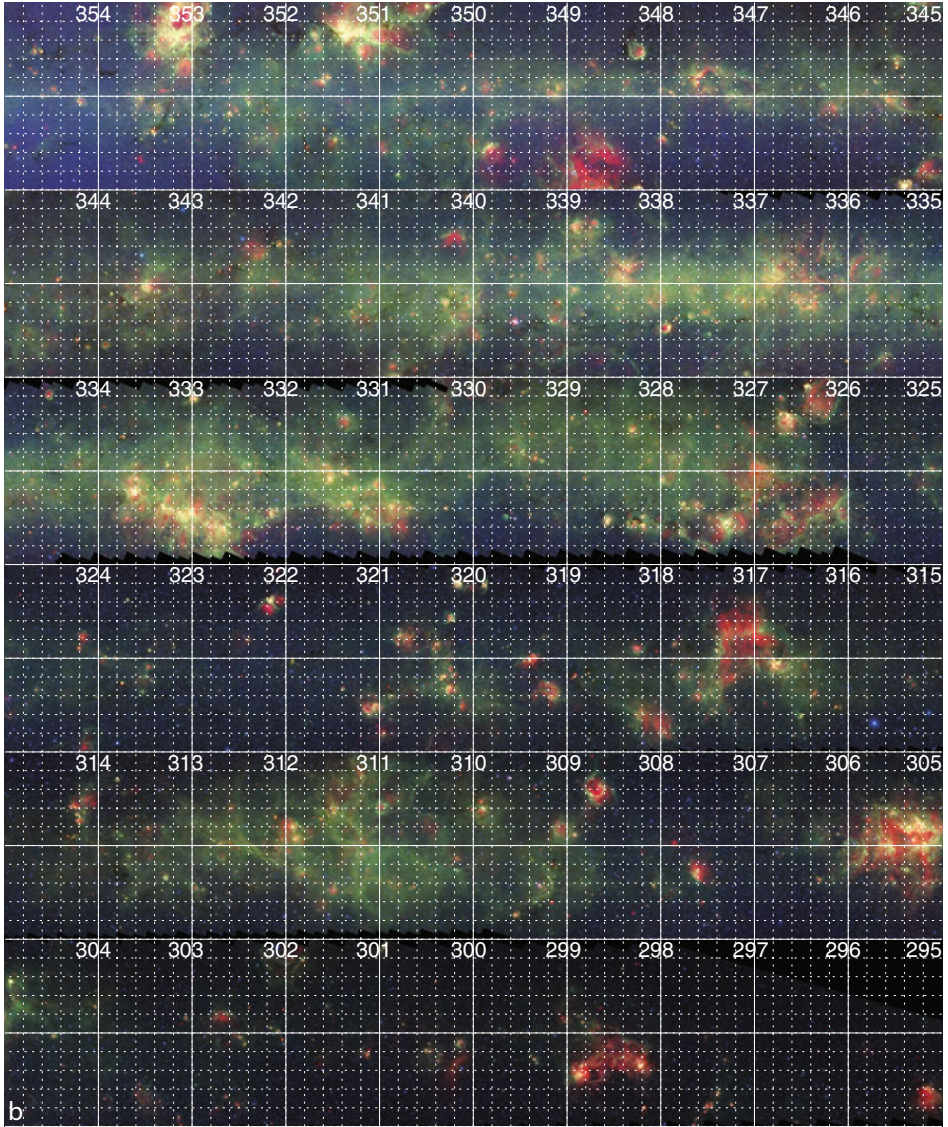
Fig. 9-10

A profile along the Galactic plane of diffuse emission in the *Spitzer*/IRAC [8.0] band from $l = 65^\circ$ to -65° averaged over five different latitude cuts. The diffuse emission is taken from a point-source subtracted image of the Galactic plane, smoothed to a resolution of $3'$. Horizontal lines at 25, 50, and 75 MJy ster $^{-1}$ are added to compare the drop in intensity with latitude. The lower envelope of diffuse emission interior to $|l| \lesssim 30^\circ$ is approximately constant, but drops much more rapidly with latitude than the diffuse emission in the longitude range $|l| \gtrsim 30^\circ$.



■ Fig. 9-11a

The GLIMPSE/MIPSGAL surveys with $4.5\ \mu\text{m}$ (blue), $8.0\ \mu\text{m}$ (green), and MIPS $24\ \mu\text{m}$ (red). Hallmarks of the mid-infrared view of the Galaxy include PAH bubbles, IRDCs, YSOs, diffuse dust/PAH emission, and millions of stars. Each strip spans 10° of the Galactic plane, centered on $l = 60^\circ$ (top panel) down to $l = 10^\circ$ (bottom strip). Many of the HII regions seen in this image, characterized by (red) $24\ \mu\text{m}$ dust emission in the inner part of the HII region and surrounded by (green) $8\ \mu\text{m}$ PAH emission in a photodissociation region, have yet to be catalogued and classified

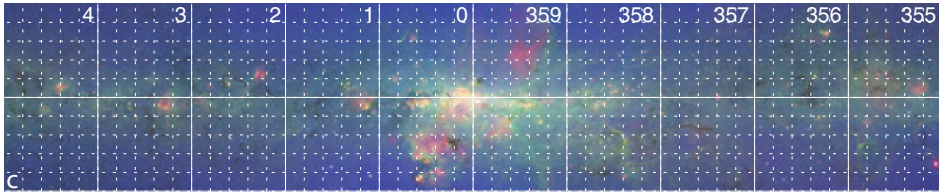


■ Fig. 9-11b

Same as ▶ Fig. 9-11a but for $l = 295 - 355^\circ$. Each strip spans 10° of the Galactic plane, centered on $l = 350^\circ$ (top panel) down to $l = 300^\circ$ (bottom strip)

3.4 Stochastic and Thermal Dust Emission

The longer wavelength emission characterized by emission in the *Spitzer*/MIPSGAL [24] band is primarily due to dust heated by starlight (Draine 2003). As shown by Draine and Li (2001), the temperature of grains smaller than about 50 \AA in a given interstellar radiation field (ISRF) strongly fluctuates with the absorption of single photons (which occurs on timescales $\geq 1.7 \text{ h}$). In contrast, grains larger than about 200 \AA have essentially a constant temperature. The former



■ Fig. 9-11c

Same as [Fig. 9-11a](#) but for $l = -5^\circ$ to 5°

process has been referred to as stochastic heating and the latter to thermal heating because it is in equilibrium with the radiation temperature of the ISRF. Thus, the $24\ \mu\text{m}$ emission measured by the MIPS GAL survey is expected to track the integrated intensity (over all wavelengths) of the ISRF convolved with the size distribution of dust grains and their spatial distribution throughout the Galaxy.

► [Figure 9-11a–c](#) show false color images of the entire GLIMPSE/MIPSGAL survey, showing the distribution of starlight in the [4.5] band (blue), PAH emission in the [8.0] band (green), and dust emission in the [24] band. The brightest $24\ \mu\text{m}$ emission shows a strikingly different morphology than the $8\ \mu\text{m}$ emission. The longer wavelength emission appears to be much patchier, brightest in the regions around hot stars (young star clusters, HII regions, young stellar objects) and AGB stars. The difference in the sky distribution of the emission from PAH and small dust grains sampled by the two bands indicate spatial variations in the ISRF, the PAH-to-dust grain ratios, or both. For example, [Draine \(2003\)](#) estimates that between 3 and $50\ \mu\text{m}$, 40% of the integrated dust emission is emitted in the wavelength range 12 – $50\ \mu\text{m}$ while 60% falls in the 3 – $12\ \mu\text{m}$ range, due to PAH emission. Both the hard component ($E > 13.6\ \text{eV}$) and the soft component ($E < 13.6\ \text{eV}$) of the ISRF are primarily produced by hot stars (O and early B stars), but since the soft UV photons can pass through HII region ionization fronts and propagate further into the ISM from hot stars than the hard UV photons, the wider diffuse distribution of $8\ \mu\text{m}$ vs the patchy $24\ \mu\text{m}$ emission is a logical consequence.

4 Star Formation

Since stars form in dense and dusty environments, it had always been anticipated that the advent of infrared astronomy would lead to a major advance in our understanding of the nature and distribution of star formation throughout the Milky Way, particularly massive star formation. What may be less appreciated is how the rapidly improving angular resolution of these surveys has led to an explosion in the number of known star formation regions at greater and greater distances from the Sun. For example, a recent *Green Bank Telescope* hydrogen recombination line survey of the Galactic plane ($l = -16^\circ$ to 67°) that targeted $24\ \mu\text{m}$ and $20\ \text{cm}$ bright diffuse sources had a 95% success rate ([Bania et al. 2010](#)). This single effort has doubled (!) the number of confirmed Galactic HII regions in this section of the Galaxy. When these new objects are plotted on a position-velocity diagram, clear evidence is seen for structure in the star-forming component of our Galaxy, including the enhanced star formation at the near end of the Long Bar, a multi-peaked radial distribution, and star formation in the Outer Arm at a distance of over $20\ \text{kpc}$ from the Sun.

It is not just the study of classical HII regions that is currently undergoing a renaissance. In a review of the stages of massive star formation, Churchwell (2002) noted that “the evolutionary stages preceding the ultra-compact HII region state are not well understood, and future efforts are likely to concentrate on these.” Infrared surveys have fulfilled this promise, yielding catalogues of entirely new classes of objects associated with star formation, e.g., infrared dark clouds, high mass stellar outflows, massive young stellar objects, and PAH bubbles. The physical properties of these objects are now being studied; their Galactic distribution has yet to be investigated.

4.1 Infrared Dark Clouds

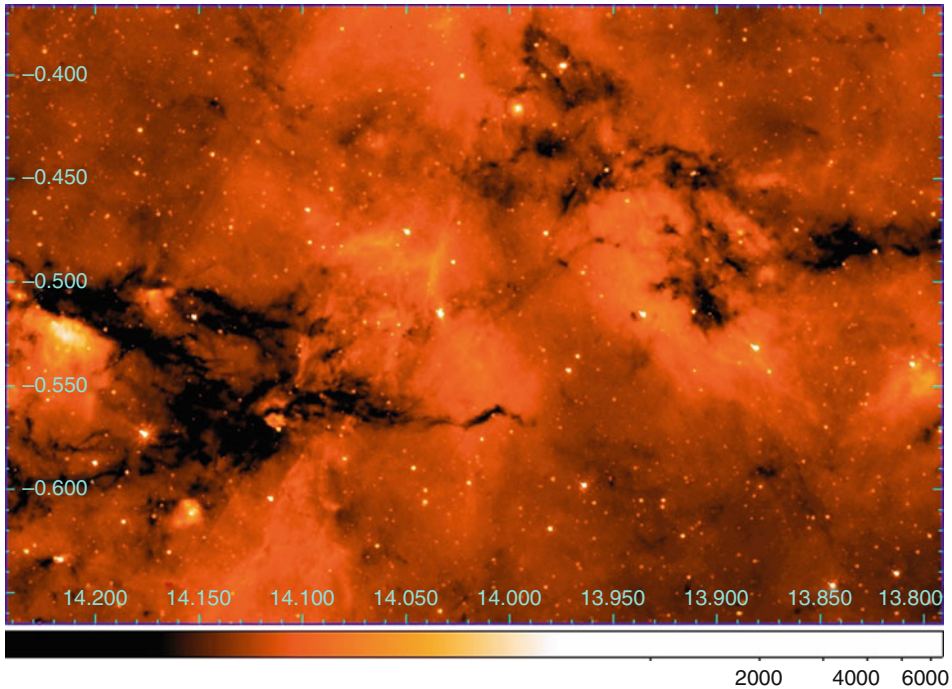
Infrared dark clouds (IRDCs) are the densest condensations in giant molecular clouds, the most likely sites of future star and cluster formation, and the repository of the Galaxy’s densest molecular constituents. IRDCs also inform us of the conditions necessary for star formation. IRDCs were first identified in the Galactic plane from *Infrared Space Observatory*/ISOCAM images, followed closely by the detection of $\sim 2,000$ IRDCs from *Midcourse Space Experiment* images. A review of much of the work done on these clouds, both in infrared and molecular emission, can be found in the review by Churchwell et al. (2009). IRDCs are typically opaque at $8\ \mu\text{m}$, implying extinctions $A_{[8.0]} > 3\ \text{mag}$ ($A_V > 70\ \text{mag}$). These objects are a striking feature of infrared surveys and can be mapped at very high angular resolution. However, the requirement of a bright diffuse mid-infrared background imposes a strong selection bias. In directions with low diffuse mid-infrared emission, i.e., increasing galactic latitude or longitude, clouds with the same physical properties as IRDCs would have to be identified by molecular line emission.

Studies with MSX and ISO first drew attention to the importance of these objects. The GLIMPSE survey, because of its greater sensitivity and spatial resolution than MSX, has revealed an even larger number of IRDCs and provided a more detailed picture of their morphologies, especially those clouds with small angular sizes. Peretto and Fuller (2009) released a catalog of 11,202 *Spitzer*/GLIMPSE IRDCs, 80% of which were not in the previous MSX-based infrared dark cloud catalogue (Simon et al. 2006).

Many studies of IRDC properties have been undertaken using various radio molecular probes and radio continuum observations. These investigations indicate typical densities $> 10^5\ \text{cm}^{-3}$, temperatures $\lesssim 20\ \text{K}$, and masses ranging from a few M_\odot to $> 10^4 M_\odot$. IRDCs are generally, but not exclusively, filamentary (see [Fig. 9-12](#)) with length-to-width ratios often well in excess of 10:1. They consist of dense condensations embedded in a lower density diffuse envelope. Ragan et al. (2009) report substructures ranging from $0.5 M_\odot$ cores to $\leq 10^4 M_\odot$ clouds,⁵ with an IRDC clump mass function whose slope becomes flatter than the slope of the Salpeter initial mass function (Salpeter 1955) for $M_{clump} < 40 M_\odot$. They suggest that this turnover may be the transition between IRDCs that produce clustered star formation and those that produce distributed star formation; further study is needed to confirm this. The same study also yielded a clump mass-radius relation $M \propto R^{2.7}$, similar to that of Williams et al. (1994).

The internal density structure of IRDC clumps and cores can be determined using mid-infrared (MIR) absorption, in addition to the standard analyses of molecular line emission or submillimeter continuum images. Abergel et al. (1998, 1996) and Bacmann et al. (1998) were the first to use extinction profiles of isolated starless cores (low-mass IRDC cores) at

⁵Bergin and Tafalla (2007) define *cores* as objects with masses $0.1\text{--}10 M_\odot$ and sizes $0.01\text{--}0.1\ \text{pc}$, *clumps* with $10\text{--}10^3 M_\odot$ and $0.1\text{--}1\ \text{pc}$, and *clouds* with $10^3\text{--}10^4 M_\odot$ and $1\text{--}10\ \text{pc}$.



■ Fig. 9-12

Example of an IRDC complex in the neighborhood of M17. This illustrates how clearly IRDCs stand out in silhouette at $8\ \mu\text{m}$ in the inner Galaxy. It also illustrates the intricate filamentary structure of IRDCs. A closer look at this complex in a three-color image shows many probable YSOs currently forming in this region. Figure from Devine (2009)

MIR wavelengths to estimate the H_2 column density profiles and total masses. They found approximately flat column density profiles out to a radius of several thousand astronomical units. Beyond this radius, the column densities sharply decrease, confirming the earlier analyses based on molecular line and submillimeter data (Andre et al. 1996; Ward-Thompson 1994). The rapidly decreasing densities beyond the central flat region indicate that pre-stellar cores are basically decoupled from their larger parent clouds, limiting the mass available to the core. The significance of this is that the clump mass function, which seems to parallel the stellar mass function but at larger masses, may depend on clump density profiles as a function of mass. This, of course, needs further independent confirmation. Additional determinations of IRDC density profiles have been obtained for several other clouds (Andre et al. 2000; Ragan et al. 2009; Ward-Thompson 1994).

Because IRDCs are dense, cold, and massive, one might expect them to be globally gravitationally unstable, but examination of many IRDCs shows that at any given time only a small fraction of the volume of a typical IRDC is involved in forming stars. Devine (2009) used the VLA to obtain high resolution, sensitive images of four IRDCs in the lines of NH_3 (1,1 and 2,2) and CCS (2_1-1_0). These images show that the NH_3 emission is generally optically thick ($\tau > 3$) and closely traces [8.0] band PAH emission, with significant velocity substructures within all

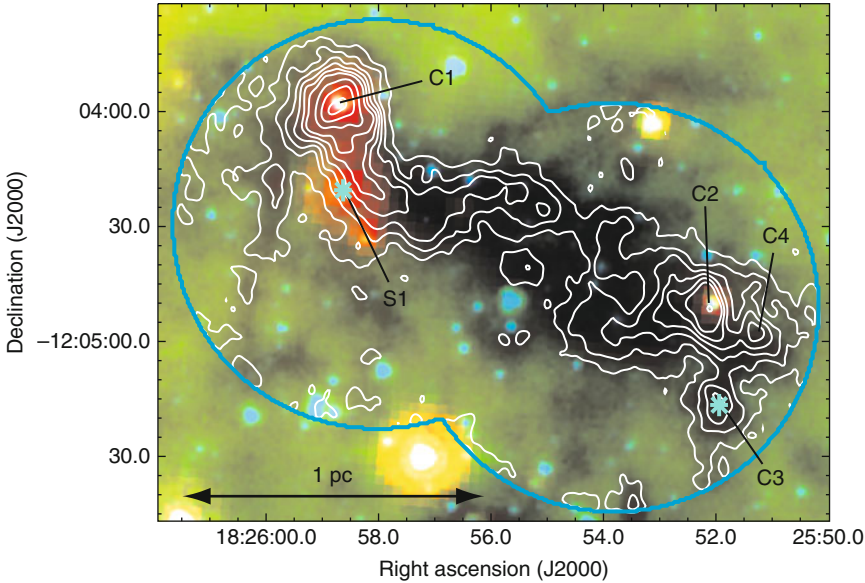


Fig. 9-13

A $\text{NH}_3(1,1)$ VLA image (contours) of the IRDC G19.30+0.07 (colored image from GLIMPSE/MIPSGAL (4.5 μm -blue, 8.0 μm -green, 24 μm -red)). The blue overlapping circles show the primary half-power beam of the VLA. The NH_3 emission closely follows the outline of the MIR dark cloud. Four dense and warm cores are indicated by labels C1-C4 from NH_3 emission, two of which are also bright 24 μm sources. The location of water masers (S1 and in C3) are shown with blue stars. C4 and S1 NH_3 sources do not stand out at MIR wavelengths, perhaps because they are too young to have heated dust such that it is bright at MIR wavelengths. Figure from Devine (2009)

four clouds. They also confirmed the general gas properties found by other investigators: densities of $\sim 10^5 \text{ cm}^{-3}$, integrated masses ranging from 1,100 to 20,000 M_\odot , gas kinetic temperatures of 15–25 K.


These observations also identified multiple hot molecular clumps in all IRDCs observed. An example of one IRDC with four hot NH_3 clumps and two H_2O masers is shown in Fig. 9-13. The NH_3 clumps, labeled C1-C4, appear to be in hydrostatic equilibrium. The clumps found in the other three IRDCs are thought to be collapsing because the mass of the NH_3 clumps is substantially larger than their virial mass inferred from line dispersions. The correspondence of some NH_3 clumps with bright MIR compact sources in IRDCs convincingly shows that these clumps are probably sites of current star formation in IRDCs.

4.2 Extended Green Objects

A new class of objects, discovered in the GLIMPSE survey, are now referred to as extended green objects (EGOs). These are diffuse sources that are bright in the [4.5] band (which is usually chosen as green in three- or four-color images using the *Spitzer*/IRAC bands). Cyganowski et al. (2008) visually identified and cataloged more than 300 EGOs in the GLIMPSE I survey.

These EGOs are found toward IRDCs and are frequently associated with bright $24\ \mu\text{m}$ sources, indicating that they are associated with an early stage of star formation.

These sources are also strongly correlated with CH_3OH (methanol) masers. Class I methanol masers (44 and 95 GHz) are collisionally excited and observationally well correlated with molecular outflows in massive star formation regions (Cragg et al. 1992; Johnston et al. 1992; Kurtz et al. 2004; Plambeck and Menten 1990), while Class II methanol masers (6.7 GHz) are radiatively pumped by IR emission from warm dust, c.f. Cragg et al. (2005) and references therein, and are exclusively associated with massive young stellar objects (Minier et al. 2003). Cyganowski et al. (2009) examined the association of EGOs with methanol masers at high spatial resolution using the VLA, finding that $\geq 64\%$ of the EGOs targeted were detected as Class II 6.7 GHz methanol masers. This maser transition is spatially concentrated ($\leq 1''$) in compact groups coincident with the center of the EGOs and $24\ \mu\text{m}$ emission.

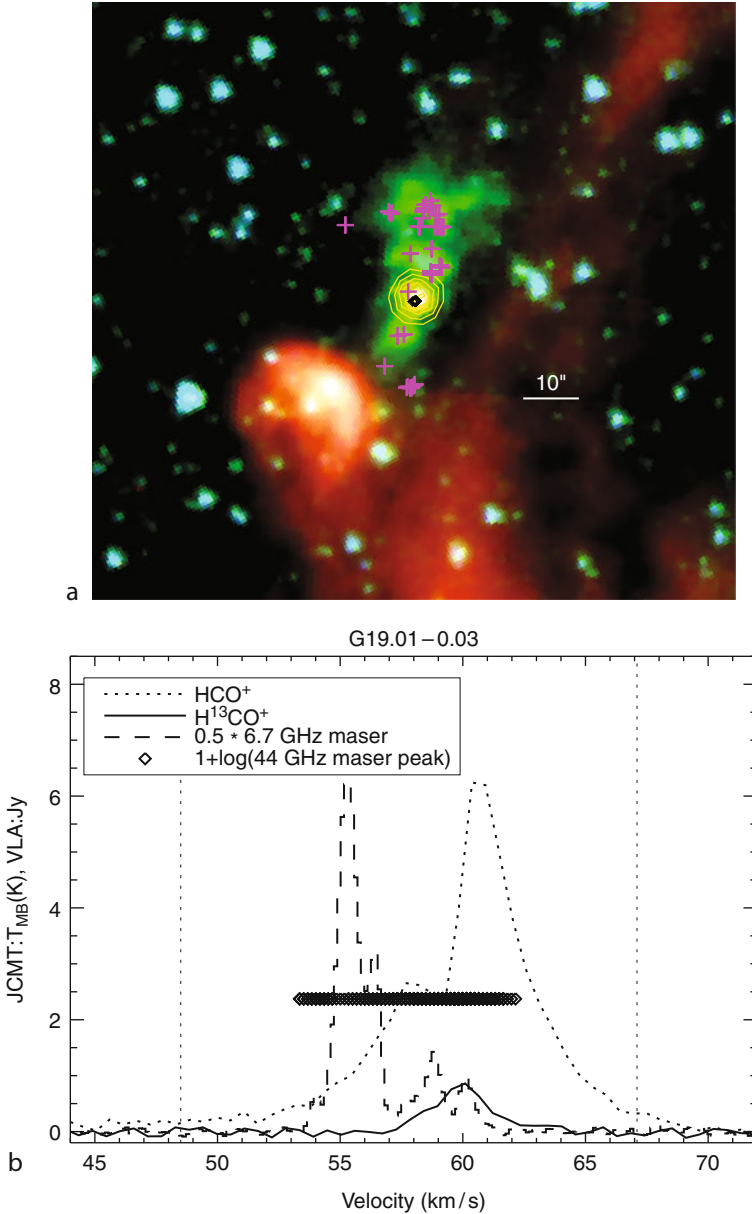
In addition, $\sim 90\%$ of the observed EGOs were also detected as Class I 44 GHz methanol masers. This emission, unlike the 6.7 GHz transition, is widely distributed over tens of arcsecs and coincides with the extended [4.5] band emission, which would be expected if this emission traces molecular outflows. The methanol emission lines are broad (typically $\geq 20\ \text{km s}^{-1}$), also consistent with outflows. The James Clerk Maxwell Telescope was also used to survey these sources in emission lines of HCO^+ (3-2) and SiO (5-4). The detection of SiO (5-4) in 90% of surveyed EGOs is consistent with ages $\sim 10^4$ years or less since SiO persists for only this long after the passage of a shock. Finally, the detection of 83% of surveyed EGOs in thermal CH_3OH ($5_{2,3} - 4_{1,3}$) emission indicates the presence of warm, dense gas. Some of these results are illustrated in  Fig. 9-14 taken from Cyganowski et al. (2009).

This work indicates that EGOs are rapidly accreting massive protostars in an early stage of formation prior to the development of hypercompact HII regions. Spectra of a few of these sources (De Buizer and Vacca 2010) indicate that the emission seen in the [4.5] band is shocked molecular hydrogen, presumably created when bipolar outflows crash into the ambient ISM. If this is generally true, mid-infrared surveys provide a very easy and quick way to identify a very early and rare stage of massive star evolution.

4.3 Massive Young Stellar Objects and the Galactic Star Formation Rate

The current census of massive stars in the Galaxy is woefully incomplete. Foreground extinction obscures many of these stars in optical and near-infrared surveys, and the youngest of these sources are still surrounded by their natal molecular and dust cocoons. Happily, mid-infrared surveys have reached the sensitivity and angular resolution that the most massive stars in the Galactic disk have mostly been *detected*. The new challenge is in *identifying* which sources, out of the hundreds of millions of objects, are the massive stars and protostars.

Since the mid-IR colors of naked OB stars are indistinguishable, the identification of these sources relies on the proximity of nearby dust. For massive stars that have had time to create HII regions, diffuse $24\ \mu\text{m}$ emission from the dusty HII regions provides a reliable signpost (Bania et al. 2010). But mid-infrared surveys are also making it possible to find massive stars at an even earlier stage of evolution. Since these massive young stellar objects (MYSOs) are still deeply embedded in their parent clouds, the large dust envelope surrounding these protostars reemits the stellar radiation at MIR wavelengths. This makes them especially bright in the $24\ \mu\text{m}$ and [8.0] bands observed by *Spitzer*.



■ Fig. 9-14

(a) A GLIMPSE/MIPSGAL image ([3.6]-blue, [4.5]-green, [8.0]-red, yellow contours-24 μm) toward the EGO G19.01-0.03. The pink and black pluses and diamond symbols represent the locations of CH₃OH 44 GHz and 6.7 GHz masers, respectively. Note the extended 4.5 μm emission and the spatial distribution of the 44 GHz CH₃OH as opposed to the central location of the 6.7 GHz maser and 24 μm dust emission. (b) The corresponding velocity range and line profiles of the CH₃OH masers (6.7 and 44 GHz), HCO⁺, and H¹³CO⁺ toward the EGO G19.01-0.03. Note the broad, self-absorbed HCO⁺ profile, the multiple, broad-velocity components of the 6.7 and 44 GHz masers. Both images are from Cyganowski et al. (2009)

A search for these objects was undertaken by Robitaille et al. (2008) who identified over 22,000 sources in the GLIMPSE I and II surveys with red mid-infrared colors, $[4.5] - [8.0] \geq 1$. Further checks on the quality of the flux densities led to a final sample of 18,949 sources. This sample is incomplete as it does not include saturated sources, sources below the sensitivity and confusion limits, and (presumably nearby) extended sources which are not included in the GLIMPSE point source catalogs. Using a combination of color-magnitude, color-color, SED, and Galactic distribution analyses of the sample of almost 19,000 intrinsically red sources, it was found that about 40% were AGB star candidates (see [Sect. 5.2](#)) and about 60% were MYSO candidates. Planetary nebulae and background galaxies together represented $\leq 2\text{--}3\%$ of the sample. Since the GLIMPSE II sample was obtained at two different epochs separated by at least 6 months, it was also possible to analyze these data for variability. About 22% of the sample was found to be variable by ≥ 0.3 mag in the [4.5] or [8.0] bands; these are likely AGB long period variables.

The identification of these massive stars in the process of formation allows for an estimate of the current star formation rate of the Galaxy. The global star formation rate (SFR) of a galaxy is a measure of its reservoir of cold, neutral gas. As a galaxy ages and its reservoir of gas is presumably depleted by trapping mass in burned out stars, its ability to create new stars also declines. The global SFR of a galaxy has implications for all the properties that determine the observed properties of a galaxy such as its integrated colors, stellar population, radiation field, and possibly even its dynamics. Robitaille and Whitney (2010) used an ensemble of YSO spectral energy distributions (Robitaille et al. 2007b) combined with a model of their spatial distribution to simulate the MYSO population detectable by GLIMPSE until it was in agreement with the observed MYSOs in the GLIMPSE catalog. They derived a global Galactic star formation rate in the range $0.7\text{--}1.5 M_{\odot} \text{ yr}^{-1}$.

This value is significantly lower than most previous estimates of the Galactic star formation rate, which are based on indirect measures of O and B stars in conjunction with an initial mass function (IMF) to extrapolate to all masses. Smith et al. (1978) found $5 M_{\odot} \text{ yr}^{-1}$ based on radio free-free emission from the Galactic disk. Diehl et al. (2006) found $4 M_{\odot} \text{ yr}^{-1}$ from the amount of ^{26}Al in the Galaxy from γ -ray flux. Misiriotis et al. (2006) found $2.7 M_{\odot} \text{ yr}^{-1}$ using the IRAS 100 μm flux and a conversion factor used for other galaxies. Murray and Rahman (2010) found $1.3 M_{\odot} \text{ yr}^{-1}$ from the total free-free emission observed by WMAP. Surprisingly, there is no review of the global star formation rate of the Galaxy that critically evaluates and compares these results.

4.3.1 PAH Bubbles and Triggered Star Formation

Bubbles are produced in the interstellar medium by the momentum and energy input from stars and stellar clusters. They are produced by asymptotic giant branch stars, planetary nebulae, supernova remnants, HII regions, and massive young stellar objects. The high angular resolution and sensitivity afforded by *Spitzer* has led to the discovery of thousands of such bubbles. Bubbles associated with stars near the end of their lifetime are discussed in [Sect. 5](#); here we summarize bubbles associated with star formation.

The most spectacular objects in the Galaxy at mid-infrared wavelengths are bubbles of diffuse emission, particularly in the [5.8] and [8.0] bands (Churchwell et al. 2007, 2006). The most luminous bubbles surround radio HII regions, presumably powered by radiation and winds from O and early B stars. About 38% of the MIR bubbles catalogued are incomplete rings and

appear to be “blown out” where the confining shell is thinnest. The bubbles have eccentricities between 0.55 and 0.85 with a peak at ~ 0.65 and are thin relative to the bubble radii. They are tightly confined to the Galactic plane with a scale height of $0.^{\circ}63 \pm 0.^{\circ}03$ (similar to that of O and B stars) and have average surface densities of $\sim 5 \text{ deg}^{-2}$ for $|l| < 10^{\circ}$ and $\geq 1.5 \text{ deg}^{-2}$ for $|l| > 10^{\circ}$.


Multiwavelength observations of these bubbles have led to new insights on the role of dust and PAHs in HII regions (Povich et al. 2007; Watson et al. 2008). The $8 \mu\text{m}$ (PAH) emission is confined to a shell that traces the photodissociation region (PDR) surrounding the HII region. The lack of $8 \mu\text{m}$ emission interior to this shell indicates that PAHs are destroyed by the stellar UV radiation. Inside the shell, the mid-infrared bubble is characterized by thermal dust emission, especially at $24 \mu\text{m}$. In some cases, there is also an $8 \mu\text{m}$ emission peak at the location of the central star, although this is probably thermal dust emission and not PAHs. Radio continuum and $24 \mu\text{m}$ emission are coincident with each other, proving that dust is present within HII regions. Both the radio continuum and $24 \mu\text{m}$ emission terminate close to the inner face of the $8 \mu\text{m}$ shell.

In HII regions dominated by stellar winds, typically those ionized by stars hotter than a spectral class O6, the immediate volume around the star is evacuated of both gas and dust producing a dip in $24 \mu\text{m}$ and radio continuum brightness, otherwise the $24 \mu\text{m}$ and radio continuum all peak at the location of the ionizing star(s). A detailed analysis of the infrared emission from the bubble N10 indicates the mid-infrared emission is due to stochastic heating, and subsequent cooling, of very small dust grains by the absorption of single UV photons (Watson et al. 2008). Everett and Churchwell (2010) have shown that the mass of dust in the N49 H2 region is very small, $\sim 0.02 M_{\odot}$, and absorbs $< 4\%$ of the stellar UV photons within the hot wind-shocked region of the nebula.

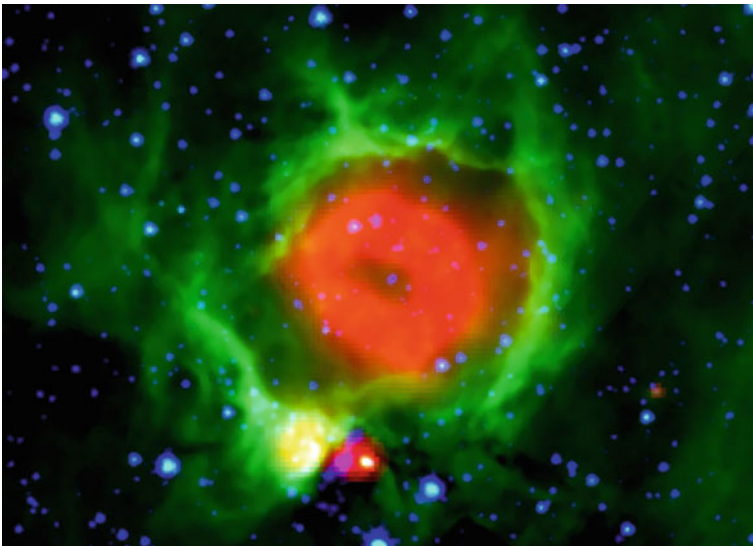
The fact that dust appears to be well mixed within the hot, wind-shocked H^+ gas has several implications (Everett and Churchwell 2010). First, HII regions are somewhat older than their dynamical ages would imply, i.e., their radii are smaller at a given age than they would be in the absence of dust. Second, the lifetime of dust in the hostile environment around hot stars must either be $\geq 10^6$ years, the minimum age of several well-studied HII regions, or else the dust must be continuously replenished. One source of dust replenishment may be the slow photo-evaporation of small dense neutral globules that were overrun by the HII region I-front and/or the release of dust from debris disks around low-mass stars in the HII region as suggested by Koenig et al. (2008). Finally, because dust in these environments is highly positively charged, dust becomes the dominant coolant. For HII regions with X-ray emitting gas, dust cooling may lead to weaker X-ray emission and lower temperatures than would be the case in the absence of dust.


Although the observed mid-infrared “bubbles” have generally been interpreted as projected three-dimensional shells, c.f. Watson et al. (2008), CO observations of a sample of these objects provide a new puzzle (Beaumont and Williams 2010). These observations show a ring of CO coincident with the the mid-infrared rings, but no evidence of CO in the center of the bubble, suggesting that these objects are true rings or tori observed face on. Beaumont and Williams (2010) infer from this that the parent molecular clouds are more sheet-like, i.e., thin in one dimension and extended in two dimensions. If this is the case, ionized gas should have a bipolar distribution with extended emission perpendicular to the thin dimension of the molecular sheet. MAGPIS 20 cm images (Helfand et al. 2006) do not indicate bipolar morphologies, but this could be because extended emission has been filtered out by the VLA. Cazzolato and Pineault (2005) also showed that expanding shells typically do not show line emission toward

their centers, even if they are surrounded by an exterior shell, due to turbulent and thermal gas motions in the shell. Further multiwavelength observations of these bubbles will be needed to resolve the question of the true bubble morphology.

A final interesting point is the relationship of these bubbles to multiple generations, or “triggered” star formation. Numerous YSOs and small secondary bubbles (presumably produced by second-generation triggered star formation) are apparent on the periphery of larger primary bubbles (presumably produced by first-generation O and B stars).  [Figure 9-15](#) shows an example of this phenomenon. YSOs were detected toward about 13% of the *Spitzer*/GLIMPSE sample of bubbles (Churchwell et al. 2007, 2006). Two mechanisms have been hypothesized for triggered star formation: (1) the Collect and Collapse (CC) model of Elmegreen and Lada (1977), and (2) the Radiation Driven Implosion (RDI) model of Bertoldi (1989). Deharveng and coworkers (see Deharveng et al. (2009) and references therein) have all identified YSOs toward numerous MIR bubbles. They find evidence for both CC and RDI formation mechanisms in compressed PDR envelopes around HII regions.

The coincidence of the PAH bubbles and these YSOs, in addition to the presence of secondary bubbles super-imposed on larger bubbles seem compelling evidence for triggered star formation. However, it is essentially impossible to prove that the YSOs would not have spontaneously formed in the absence of the HII region. Even if the YSOs on the peripheries of HII



 Fig. 9-15

An image of the MIR wind-dominated bubble N49 composed of GLIMPSE/MIPSGAL images from Churchwell et al. (2006) with $4.5\ \mu\text{m}$ (blue), $8.0\ \mu\text{m}$ (green), and $24\ \mu\text{m}$ (red). An O5 V star is located in a dip at the center. A bright $24\ \mu\text{m}$ shell of thermal dust emission surrounds the central dip which is within the hot wind-shocked, ionized gas. A PDR shell traced by $8\ \mu\text{m}$ emission surrounds the H2 region. Along the bottom of the PDR shell there are three YSOs; the right most is bright at $24\ \mu\text{m}$, the middle one is an EGO with extended emission at $4.5\ \mu\text{m}$ and is also bright at $24\ \mu\text{m}$, the left most with the yellow color is a compact H2 region and is bright at 24 and $8\ \mu\text{m}$. These suggest triggered star formation

regions were triggered by the HII region, only 10–15% of bubbles show probable evidence of triggering (Churchwell et al. 2006) suggesting that this mechanism, although important for astrophysical reasons, may not be the dominant star formation mechanism.

4.4 Massive Star Formation Regions: A Case Study

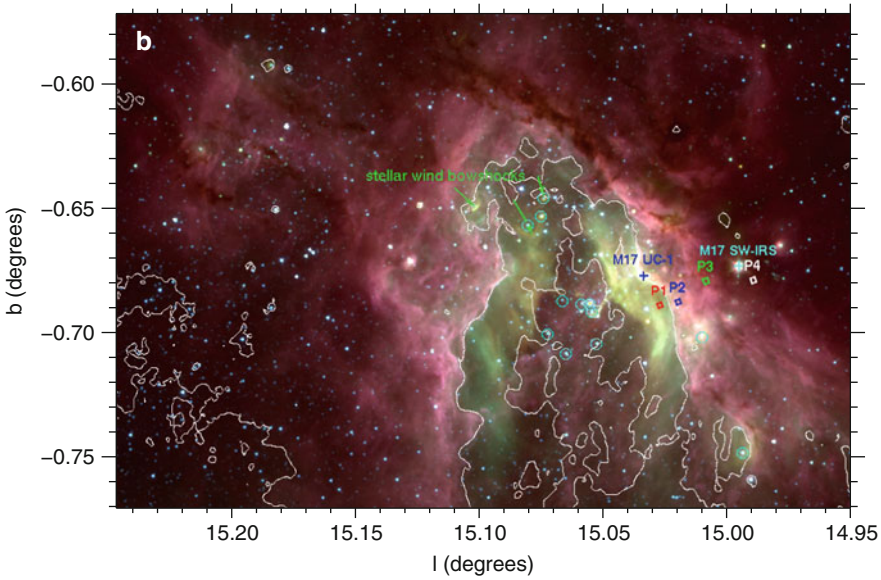
Massive star formation regions (MSFRs) are rare but important. They produce the bulk of the radiant energy in the Galaxy, are responsible for heavy element enrichment of the ISM, impact the dynamics of the ISM in their neighborhoods via stellar winds and radiation pressure, heat their natal molecular clouds, and drive complex exothermic chemical reactions. They are also very prominent at mid-infrared wavelengths. Due to their intense UV photon fluxes and strong stellar winds, MSFRs generally have a central wind-evacuated region surrounded by a shocked, very hot plasma outside of which is a photodissociated (PDR) shell. These are the stars responsible for producing the bubbles discussed in the previous section and for ultimately dispersing their natal molecular cloud.

For many years, our understanding of massive star formation regions relied on detailed investigation of the Orion nebula which is close enough that extinction (in the optical), or limited angular resolution, and sensitivity (in the infrared) were not insurmountable obstacles. With current instrumentation, detailed investigations of the stellar and interstellar content of hundreds of MSFRs are now possible. Here we summarize what has been learned for one nearby example, M17.

M17 is among the most massive and youngest star formation regions within 2 kpc of the Sun. Povich et al. (2007) present a recent analysis of the spectral energy distribution of M17 as a function of wavelength from $\sim 1 \mu\text{m}$ to 90 cm. In particular, they identified changes in the SEDs corresponding to additional emission in the *Spitzer*/IRAC [3.6], [5.8], and [8.0] bands beyond what would be expected from thermal dust emission. This excess emission is produced by PAHs. This work also presented evidence that PAHs are destroyed in the M17 HII region by characterizing the spatial variation of the *Spitzer*/IRAC [4.5] band (which contains no PAH feature) to the other bands (► Fig. 9-16). The destruction of PAHs in the region of ionized gas was confirmed with *Spitzer*/IRS spectra taken at four locations (► Fig. 9-17). These spectra showed a rapid drop in PAH emission as one moves from the PDR region across the M17 SW arm into the HII region.

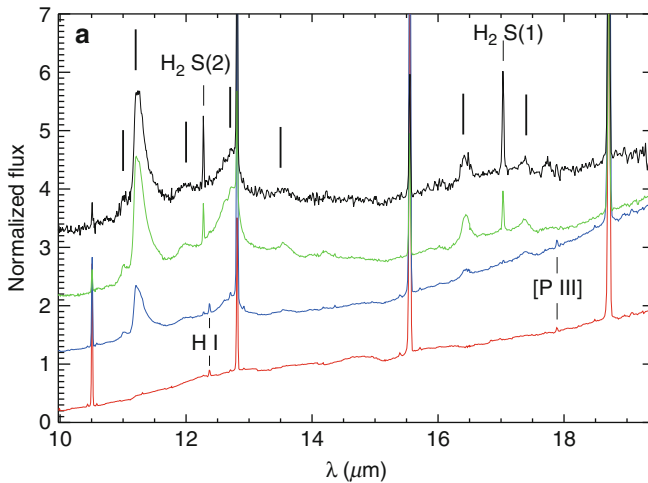
This study also showed the presence of hot, diffuse, soft X-ray emitting gas in the central cavity between the SW and N arms of M17 (► Fig. 9-18). The hot X-ray emitting gas appears to be flowing unopposed from the central cavity where the two arms of M17 open outward. This illustrates the expected general morphology of MSFRs (and wind-shocked HII regions) that contain multiple massive stars with strong winds, namely, a hot, wind-shocked X-ray emitting central cavity, surrounded by a shell of warm thermally emitting dust that is bright at 20–30 μm and lies inside the cooler $\sim 10^4$ K ionized hydrogen traced by radio continuum emission. The PDR traced by PAH emission (IRAC bands [5.8] and especially [8.0]) forms an envelope around both the north and southwest arms of M17.

The more recent study of Povich et al. (2009) puts M17 into an even larger context (► Fig. 9-19), finding a mid-infrared bubble, M17EB, with a diameter of ~ 0.95 (17.5 pc at a distance of 2 kpc). This bubble is bounded on one side by the M17 HII region and the other side by the molecular cloud, MC G15.9-0.7. Star formation is also present where the M17EB bubble impinges on this molecular cloud. M17EB and MC G15.9-0.7 are seen in CO (2-1) emission



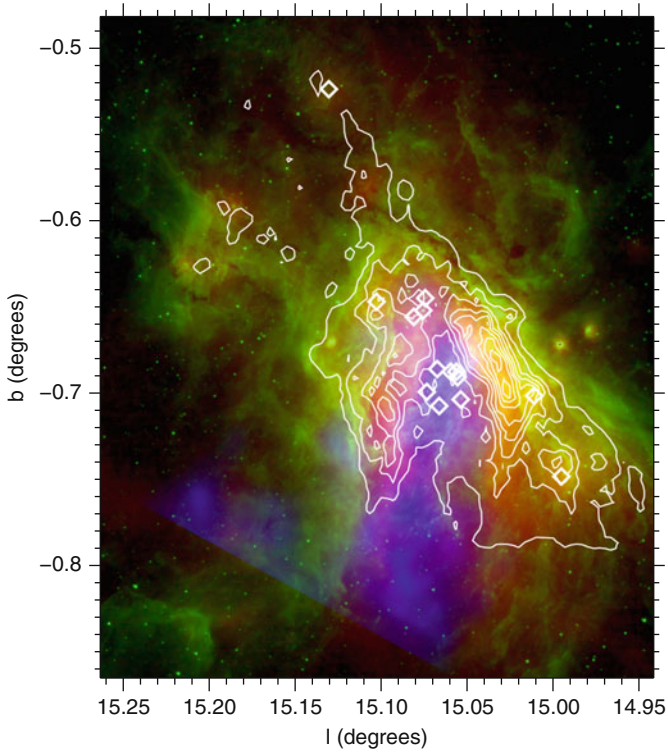
■ Fig. 9-16

M17 in IRAC bands [3.6]-blue, [4.5]-green, and [5.8]-red. O stars are indicated by cyan circles, wind-blown bow shocks are indicated by green arrows, and the boxes labeled *P1-P4* show the locations of IRS spectra, using the same color coding as the spectra in the right figure. *P1* probes the H₂ region; *P2* & *P3* probe the PDR region along the M17 SW arm, and *P4* probes the M17SW molecular cloud. The white contours show the region inside of which PAHs have been destroyed, as determined from the ratio of [5.8]/[4.5]. Figure from Povich et al. (2007)



■ Fig. 9-17

Spitzer/IRS spectra at locations *P1-P4* shown in Fig. 9-16. The spectra have been vertically shifted for clarity. Note the spectral change from nebular fine structure lines at *P1*, to increasing intensity of PAHs at *P2* and *P3*, to an increasing intensity of H₂ lines from *P2* to *P4*. Figure from Povich et al. (2007)

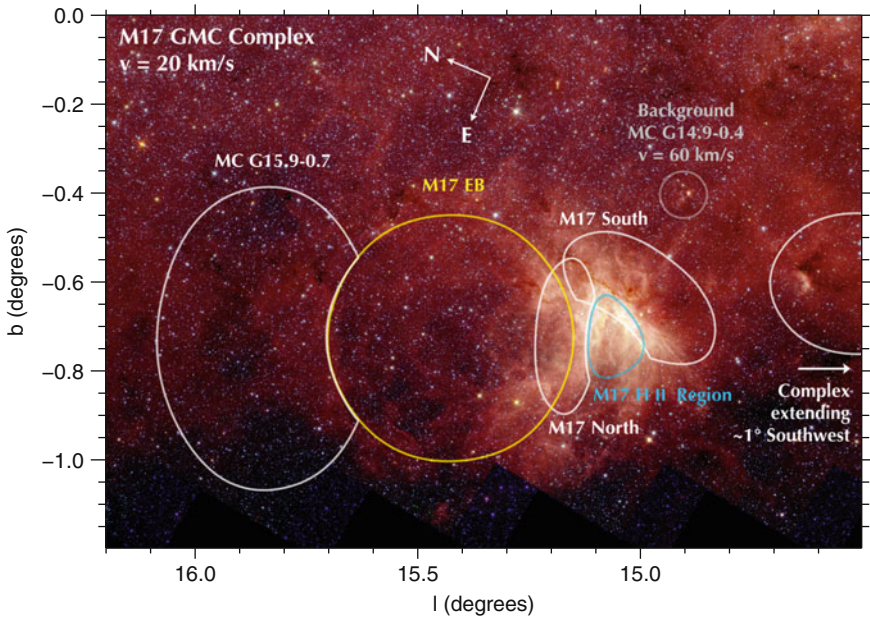


■ Fig. 9-18

A composite color image of M17 showing soft X-ray diffuse emission (0.5–2.0 keV) in blue, thermal dust emission at 21.3 μm MSX emission in red, and PAH emission from IRAC [5.8] in green. The white contours show 20 cm continuum emission indicating the distribution of $\sim 10^4$ K ionized gas. Image from Povich et al. (2007)

(► Fig. 9-20) at the same velocity as M17 ($\sim 19 \text{ km s}^{-1}$). The CO data indicate a gas mass of $\sim 1.4 \times 10^5 M_{\odot}$ in the velocity range 12–26 km s^{-1} .

YSO candidates were identified throughout the entire M17 complex using both MIR color-color analyses and the spectral energy distribution (SED) analyses using the Monte Carlo radiative transfer models of massive YSOs developed by Whitney et al. (2004) combined with the model-fitting routine of Robitaille et al. (2007b). A control field was used to estimate the fraction of contamination by unassociated YSOs, which in this part of the Galactic plane is about 50%! Five candidate ionizing stars of M17EB were identified near the center of M17EB with an estimated age of 2–5 Myr, compared to ~ 0.5 Myr for stars with mass $\geq 3M_{\odot}$ in the ionizing cluster of M17 (NGC6618). The YSO populations are concentrated in M17 itself, along the periphery of M17EB, near the center of M17EB, and in MC G15.9-0.7. Based on relative ages and locations, it seems plausible that NGC6618, the ionizing cluster of M17, may have been triggered by the older cluster responsible for the creation and ionization of M17EB. The expansion of M17EB may also be triggering current star formation in MC G15.9-0.7. It was also



■ Fig. 9-19

A schematic diagram showing the approximate locations and boundaries of several large-scale features associated with the M17 H₂ region superimposed on a GLIMPSE [5.8]-red, [4.5]-green, and [3.6]-blue image (Povich et al. 2009). Labeled features are discussed in the text

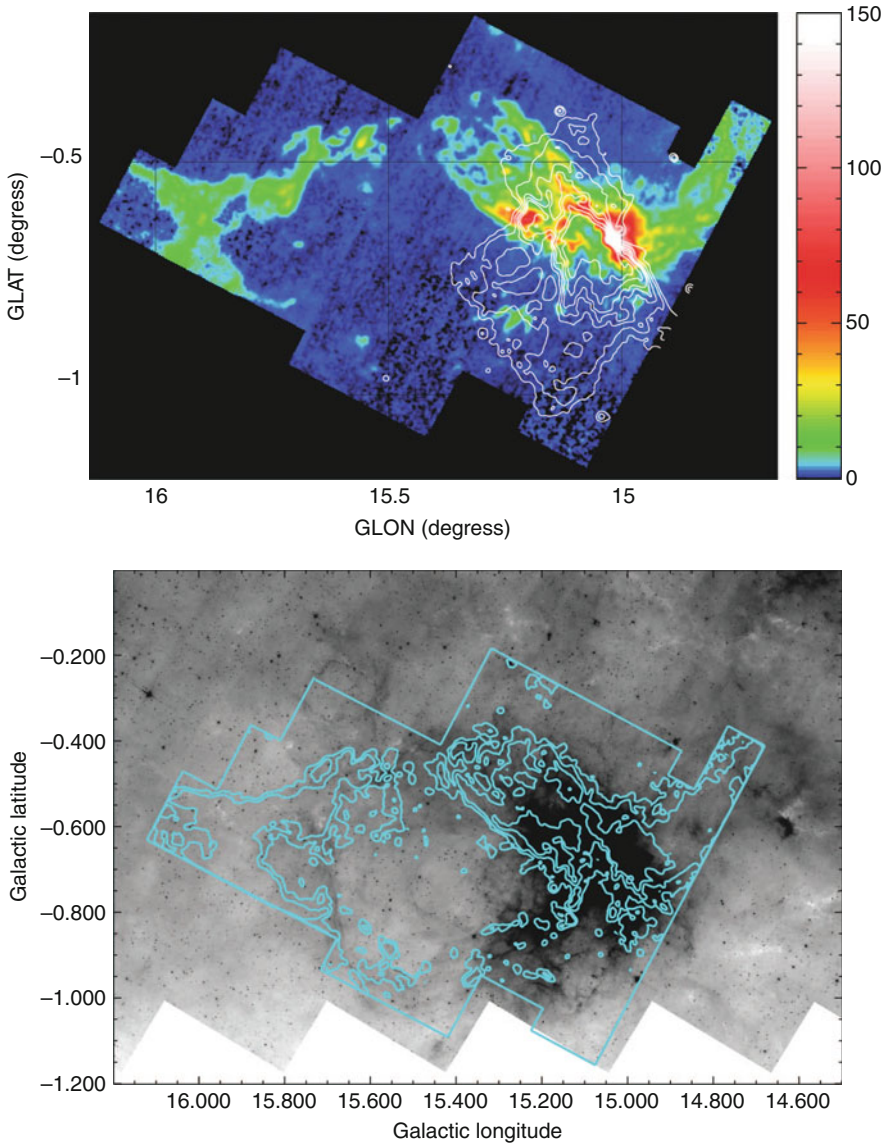
noted that essentially all the O stars in NGC6618 are likely to be binary systems for consistency with the inferred UV photon flux and current distance estimates to M17.

5 Evolved Stars

In the course of their evolution nearly all stars will lose mass, some of which will condense into dust. Infrared observations are thus important for determining the effect of this mass loss on stellar evolution and constraining the return of processed material to the interstellar medium. We review some of the results working our way upward in mass, but starting with a brief summary of stellar variability in the infrared.

5.1 Variable Stars

At some level all stars are variable, but stars with periodic variability or massive stars with significant mass loss have been important in the history of astronomy. Cepheid variables are among the key standard candles in setting the extragalactic distance scale. RR Lyrae stars were used for one of the first determinations of the distance to the center of the Galaxy. Evolved massive stars are frequently large amplitude and long period variables; study of these sources are key



■ Fig. 9-20

[*Top*]. A $^{13}\text{CO}(1-0)$ image of M17 and part of the M17EB shell (color with $T_A \Delta v (12-26 \text{ km s}^{-1})$ in K km s^{-1} shown in the scale bar at right) obtained with the Heinrich Hertz Telescope (HHT). The contours are 90 cm VLA observations at intervals of 5, 10, 20, 30, 40, 60, and 80% of peak, which traces H^+ emission. [*Bottom*] The HHT $^{12}\text{CO}(1-0)$ emission at 19 km s^{-1} superimposed on the GLIMPSE [8.0] image. The large M17EB molecular shell is clearly apparent in CO as well as at $8 \mu\text{m}$ emission. Figure from Povich et al. (2009)

for understanding the role of mass loss in stellar evolution as well as constraining the return of chemically enriched material into the interstellar medium. The variability associated with mass transfer and eclipses in binary systems provides important constraints on models of stellar evolution. Several programs are dedicated to searching for microlensing, but have also turned up large numbers of other variable sources in abundance.


With the exception of the work on evolved stars, most of the study of variable stars has been the domain of optical astronomy. Large-scale deep surveys, e.g., the Optical Gravitational Lensing Experiment (OGLE), have produced a catalog of 200,000 potentially variable sources (Wozniak et al. 2002) in selected low-extinction directions, like Baades window. But it is often noted that the two principal difficulties using optical surveys of variable stars to probe the structure of the Galaxy have been extinction and lack of uniform coverage (Paczynski 1997).

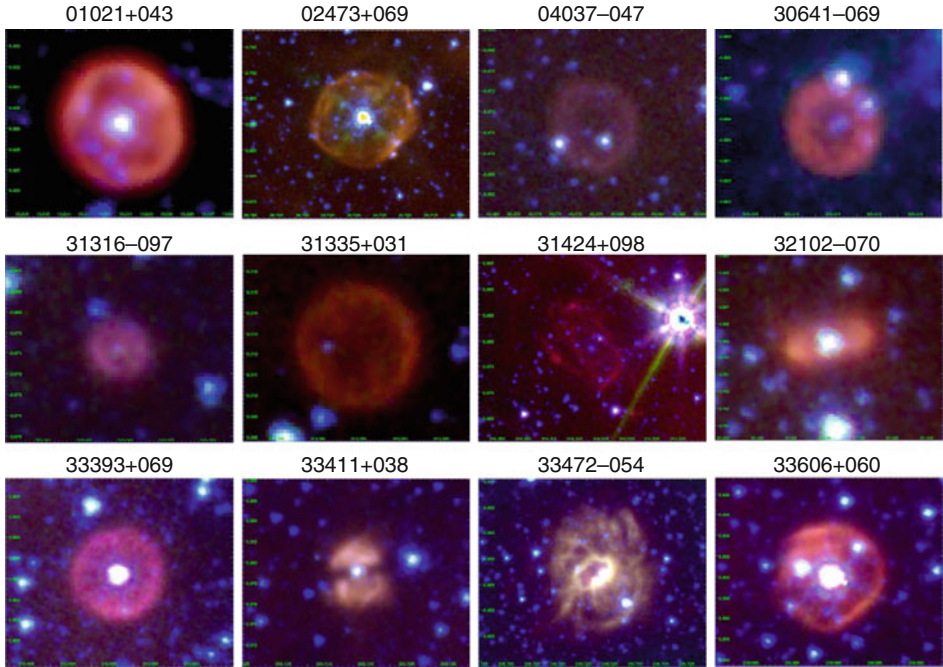
Surprisingly, only the ongoing near-infrared VVV survey (Minniti et al. 2010) has been designed to produce the light curves needed to identify different classes of variable sources. Most previous surveys, with only a few epochs of observation, only allow for variable sources to be flagged. These studies tend to preferentially select large-amplitude or long-period variables. For example, Robitaille et al. (2007a) used the MSX and GLIMPSE/MIPSGAL surveys to identify candidate mid-infrared variables by comparing flux densities at 8.0 vs 8.28 μm and secondarily at 24 vs 21.3 μm . By requiring that the amplitude of variation exceed a factor of two, the number of candidate-variable sources was winnowed from almost 52,000 sources to 592, most of which are expected to be asymptotic giant branch stars.

5.2 Asymptotic Giant Stars

During the asymptotic giant branch (AGB) phase of low and intermediate mass stars, stars develop dusty circumstellar shells. These stars provide a significant source of carbon and oxygen and dust to the interstellar medium; an extensive review can be found in Habing (1996). A survey of the LMC (Srinivasan et al. 2009), which has the advantage that the sources are at a common distance and extinction, used infrared SEDs to identify different classes of AGB stars, finding that the dust mass returned to the LMC from these sources is $6\text{--}13 \times 10^{-3} M_{\odot} \text{ yr}^{-1}$. Using a catalog of red GLIMPSE objects, Robitaille et al. (2008) compared the observed Galactic distribution to a synthetic distribution, assuming that Galactic AGB stars have similar infrared colors to those in the LMC. They estimate that 30–50% of the 22,000 red sources are AGB stars. Much more work characterizing the galactic population of AGB stars remains to be done.

5.3 Planetary Nebulae

The mid-infrared colors of planetary nebulae strongly overlap with HII regions and young stellar objects (YSOs). The identification of PN in the mid-infrared therefore requires confirming images or spectra. In  Fig. 9-21, a montage of PNe images from the GLIMPSE survey are shown to illustrate some of the morphologies seen in the IRAC bands. Kwok et al. (2008) analyzed the SEDs for 30 previously identified PNe in the GLIMPSE survey. Phillips and Ramos-Larios (2008b) used 2MASS and GLIMPSE, finding that although their colors are very similar to HII regions for identification purposes, their intensity profiles have broad wings that increase with wavelength which they attribute to a photodissociation region (PDR) around PNe. Using this, they identified a set of candidate PNe (Phillips and Ramos-Larios 2008a) that await



■ Fig. 9-21

A montage of several planetary nebulae seen in the GLIMPSE survey. Colors are [8.0]-red, [4.5]-green, and [3.6]-blue

spectroscopic confirmation. Hora et al. (2006, 2004) report *Spitzer*/IRAC observations of several PNe, including the Helix nebula. They found these objects are quite red, $[3.6] - [4.5] > 0.6$ and $[5.8] - [8.0] > 1.0$, whereas the central stars are ~ 0 in both colors. They also suggest that the [8.0] band may have significant contributions from H_2 and [ArIII] line emission.

PNe have also been imaged with *Spitzer*/MIPS. Observations of NGC2346 (Su et al. 2004) show an edge-on toroid at $70 \mu\text{m}$, the tips of the bipolar outflow at $24 \mu\text{m}$, and indicate that dust resides in the ionized nebula. The Helix nebula (Su et al. 2007) shows a debris disk of dust mass $\sim 0.13 M_\odot$ around the central star based on the observed SED. The $24 \mu\text{m}$ images also show very bright emission surrounding the central star, indicating dust close to the star. This they interpret as further evidence for a debris disk. Their IRS spectrum also shows a bright [OIV] $25.9 \mu\text{m}$ emission line contribution to the $24 \mu\text{m}$ band. NGC 650 also appears to have a nearly edge-on torus (Ueta 2006) as well as $24 \mu\text{m}$ emission around the central star. As more far-infrared data become available, more new insights on the nature of PNe are expected.

5.4 Luminous Blue Variables and Wolf-Rayet Stars

One of the surprises of the MIPS GAL survey was the presence of (at least) 416 disk and ring sources seen in $24 \mu\text{m}$ (Mizuno et al. 2010), which break into 54 objects with central sources, 112 rings, 226 disks, 24 two-lobed objects, two filamentary, and 10 miscellaneous objects.

Approximately 80% of these objects have no $8\ \mu\text{m}$ counterpart indicating that most of them are probably not PAH bubbles or standard HII regions. Further work (Gvaramadze et al. 2010; Mizuno et al. 2010; Wachter et al. 2010) has shown that these sources are a heterogeneous set, including some planetary nebulae and circumstellar envelopes around giant and supergiant stars.

However, a significant fraction of these sources are associated with various stages of massive star evolution. Near- and mid-infrared studies, c.f. Mauerhan et al. (2009) and references therein, have increased the sample of known Wolf-Rayet stars by 30%. When surveys and spectroscopic follow-up are complete, the sample is likely to more than double. Some of the sources associated with the $24\ \mu\text{m}$ shells have been found to be Be stars and Wolf-Rayet stars. Most significantly, several shells are associated with Luminous Blue Variables (LBVs). These stars are thought to be a short-lived stage of massive star evolution, preceding the Wolf-Rayet phase, during which the star loses a significant fraction of its mass. Studies of these phases of massive star evolution have always been dogged by small sample sizes. With these new infrared-selected candidates, that may soon change.

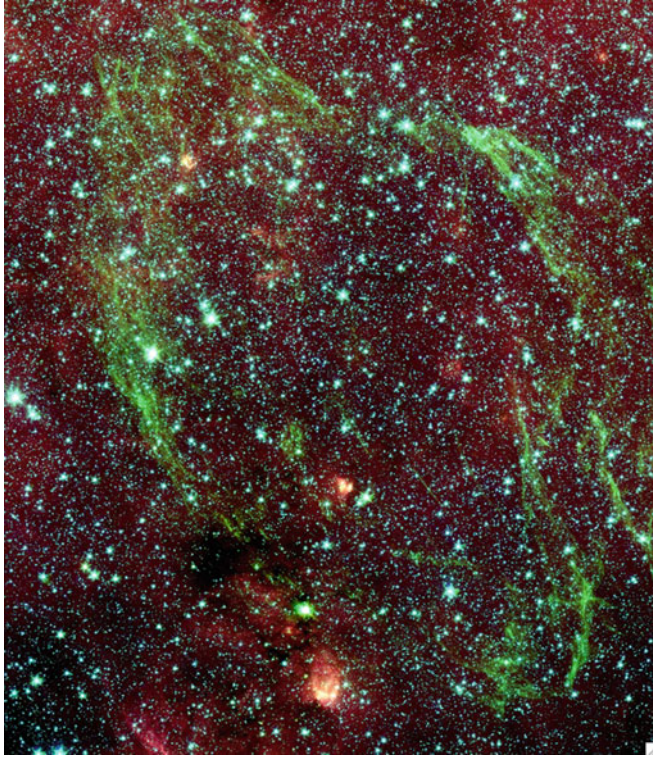
5.5 Supernova Remnants

Supernova remnants (SNRs) generally do not stand out in the mid-infrared, unless they occur in the vicinity of molecular gas. Reach et al. (2006) visually searched the GLIMPSE images for all 95 known SNRs that lie in the survey area. Only 18 (19%) of the SNRs were unambiguously detected. The low detection rate may have several causes: confusion with HII regions, shock velocities that do not produce emission lines in the *Spitzer*/IRAC bands, and the contamination of mid-infrared diffuse emission with SNR shell emission. Some, but not all, of the 18 SNRs are traced by excess emission in the [4.5] band due to lines of shocked molecular hydrogen. This includes W44 (🔗 Fig. 9-22), G311.5-0.3, and RCW103. This work also shows how SNRs can produce a wide range of mid-infrared molecular and atomic line emission. The variation in these emission lines from object to object, and even within a single object, mean that SNRs cannot be isolated to a small color-color space at mid-infrared wavelengths.

6 Limitations and Lessons Learned

Before the opening of the infrared frontier, the study of stars, star formation, and interstellar medium was the domain of the optical and radio astronomers. Looking back on what infrared investigations have revealed so far, it is instructive to reflect on where these previous eras were limited, and where new opportunities may lie.

The unbreachable limitation on optical investigations of the Galaxy is the patchy extinction principally, but not exclusively, found along the plane of the Galaxy. As a result, most of the inner Milky Way will forever be off-limits to the optical astronomer. This limitation affects much of the triaxial bulge and also means that all of the Galactic disk more distant than a few kiloparsecs is unreachable. As new optical surveys come on line to probe the time-variation, parallax, proper motions, radial velocities, and metallicities of stellar sources, it will be important to remember this limitation.



■ Fig. 9-22

GLIMPSE image of W44, one of the few supernova remnants to show clear mid-infrared emission, due to emission from shocked molecular hydrogen. See Reach et al. (2006) for more details

In contrast, radio surveys of the Galaxy, which suffer no extinction, have been limited principally by angular resolution. Until very recently with VLA 6 cm CORNISH Galactic Plane Survey (Purcell et al. 2008), there have been no radio surveys of the Galactic plane that come close to matching the \sim arcsec resolution available to optical and infrared surveys. This limitation also produces a distance bias in Galactic investigations, since distant clouds and star formation regions are unresolved. As a result, most of the focus has been on the more nearby resolved structures.

The combination of these two selection effects, extinction at optical wavelengths and angular size for radio wavelength surveys, means that beyond a few kiloparsecs from the Sun our knowledge of the Galaxy is grossly incomplete. In his introduction to the two-volume, 1,900-page handbook of star formation regions in the Galaxy, Reipurth (2008) notes that all of the star formation regions described in the volumes lie within ~ 2 kpc of the Sun. Students looking to get into Galactic astronomy should take heart. Most of the Galaxy yet remains to be explored!

Although infrared studies of the Galaxy have the advantage of low extinction and high angular resolution (and sensitivity), they too have limitations that must be borne in mind. The flip side of low extinction and a large number of sources is source confusion. Confusion imposes a brightness limitation on stellar catalogues that will vary with direction. Moreover, the current

generation of infrared surveys are all photometric. Very few velocity-resolved infrared *spectral* surveys, either for the stellar component or the diffuse emission, are on the horizon. This means that velocities, and therefore kinematic distances, require follow-up at other wavelengths.

Finally, a limitation that has dogged all Galactic investigations is the issue of sky coverage. In order to draw conclusions about the structure of the whole Galaxy it is wise to survey the entire Galaxy (or at the very least the full Galactic plane). This is particularly challenging for ground-based studies as it requires two observing sites. In this respect, infrared Galactic astronomy has been fortunate as it has come of age at the same time that data acquisition and analysis software are up to the task. But synthesis of all of this data into useful knowledge will remain a challenge for us all. Galactic astronomy has a lot of trees to distract us. But we are definitely making progress on understanding the forest.

Acknowledgments

The authors would like to acknowledge the support and tireless effort of the entire GLIMPSE team, particularly Brian Babler, Marilyn Meade, Barbara Whitney, Thomas Robitaille, Remy Indebetouw, Christer Watson, Matt Povich, Claudia Cyganowski, Katie Devine, and Stephan Jansen. They would also like to thank Robert Hurt, Sean Carey, and the members of the MIPS-GAL team for their work in producing Fig. 11. This work has made extensive use of the NASA Astrophysical Data System and the SIMBAD database, operated at CDS, Strasbourg, France. EBC would like to acknowledge partial support from *Spitzer*/NASA/JPL contracts 1224653, 1298148, 1347278, 1368699, and NSF grant AST-0808119 to the University of Wisconsin-Madison. RAB would like to acknowledge the support of *Spitzer*/NASA/JPL contract 1368014 to the University of Wisconsin-Whitewater.

Cross-References

- ▶ [Astrophysics of Galactic Charged Cosmic Rays](#)
- ▶ [Dynamics of Disks and Warps](#)
- ▶ [Galactic Distance Scales](#)
- ▶ [Galactic Neutral Hydrogen](#)
- ▶ [High-Velocity Clouds](#)
- ▶ [Interstellar PAHs and Dust](#)
- ▶ [Magnetic Fields in Galaxies](#)
- ▶ [Mass Distribution and Rotation Curve in the Galaxy](#)

References

- | | |
|--|--|
| <p>Abergel, A., Bernard, J. P., Boulanger, F., Cesarsky, C., Desert, F. X., Falgarone, E., Lagache, G., Perault, M., Puget, J.-L., Reach, W. T., Nordh, L., Olofsson, G., Hultdtgren, M., Kaas, A. A., Andre, P., Bontemps, S., Burgdorf, M., Copet, E., Davies, J., Montmerle, T., Persi, P., & Sibille, F. 1996, <i>A&A</i>, 315, L329</p> | <p>Abergel, A., Bernard, J. P., Boulanger, F., Desert, F. X., Lagache, G., Puget, J. L., Reach, W. T., Falgarone, E., Nordh, L., Olofsson, G., Andre, P., Bacmann, A., & Ristorcelli, I. 1998, <i>Star Formation with the Infrared Space Observatory</i>, 132, 220</p> <p>Alard, C. 2001, <i>A&A</i>, 379, L44</p> |
|--|--|

- Alexander, M. J., Kobulnicky, H. A., Clemens, D. P., Jameson, K., Pinnick, A., & Pavel, M. 2009, *AJ*, 137, 4824
- Alves, D. R. 2000, *ApJ*, 539, 732
- Andre, P., Ward-Thompson, D., & Barsony, M. 2000, *Protostars and Planets IV*, eds. V. Mannings, A. P. Boss, S. S. Russell (Tucson: University of Arizona Press), 59
- Andre, P., Ward-Thompson, D., & Motte, F. 1996, *A&A*, 314, 625
- Athanassoula, E. 2007, *MNRAS*, 377, 1569
- Babusiaux, C., Gomez, A., Hill, V., Royer, F., Zoccali, M., Arenou, F., Fux, R., Lecureur, A., Schultheis, M., Barbuy, B., Minniti, D., & Ortolani, S. 2010, *A&A*, 519, 77
- Bacmann, A., Andre, P., Abergel, A., Bernard, J. P., Puget, J. L., Bontemps, S., & Ward-Thompson, D. 1998, *Star Formation with the Infrared Space Observatory*, 132, 307
- Bakos, J., Trujillo, I., & Pohlen, M. 2008, *ApJ*, 683, L103
- Bania, T. M., Anderson, L. D., Balsler, D. S., & Rood, R. T. 2010, *ApJL*, 718, L106
- Beaumont, C. N., & Williams, J. P. 2010, *ApJ*, 709, 791
- Beichman, C. A., Neugebauer, G., Habing, H. J., Clegg, P. E., & Chester, T. J. 1988, *Infrared astronomical satellite (IRAS) catalogs and atlases*, 1, 1
- Benjamin, R. A., Churchwell, E., Babler, B. L., Bania, T. M., Clemens, D. P., Cohen, M., Dickey, J. M., Indebetouw, R., Jackson, J. M., Kobulnicky, H. A., Lazarian, A., Marston, A. P., Mathis, J. S., Meade, M. R., Seager, S., Stolovy, S. R., Watson, C., Whitney, B. A., Wolff, M. J., & Wolfire, M. G. 2003, *PASP*, 115, 953
- Benjamin, R. A., Churchwell, E., Babler, B. L., Indebetouw, R., Meade, M. R., Whitney, B. A., Watson, C., Wolfire, M. G., Wolff, M. J., Ignace, R., Bania, T. M., Bracker, S., Clemens, D. P., Chomiuk, L., Cohen, M., Dickey, J. M., Jackson, J. M., Kobulnicky, H. A., Mercer, E. P., Mathis, J. S., Stolovy, S. R., & Uzpen, B. 2005, *ApJ*, 630, L149
- Bergin, E. A., & Tafalla, M. 2007, *ARAA*, 45, 339
- Bertoldi, F. 1989, *ApJ*, 346, 735
- Binney, J., Gerhard, O. E., Stark, A. A., Bally, J., & Uchida, K. I. 1991, *MNRAS*, 252, 210
- Binney, J., & Merrifield, M. 1998, *Galactic Astronomy* (Princeton: Princeton University Press)
- Binney, J., & Tremaine, S. 2008, *Galactic Dynamics* (2nd ed.; Princeton: Princeton University Press)
- Bissantz, N., Englmaier, P., & Gerhard, O. 2003, *MNRAS*, 340, 949
- Blitz, L., & Spiegel, D. N. 1991, *ApJ*, 379, 631
- Block, D. L., Buta, R., Knapen, J. H., Elmegreen, D. M., Elmegreen, B. G., & Puerari, I. 2004, *AJ*, 128, 183
- Block, D. L., & Wainscoat, R. J. 1991, *Nature*, 353, 48
- Boggess, N. W., Mather, J. C., Weiss, R., Bennett, C. L., Cheng, E. S., Dwek, E., Gulkis, S., Hauser, M. G., Janssen, M. A., Kelsall, T., Meyer, S. S., Moseley, S. H., Murdock, T. L., Shafer, R. A., Silverberg, R. F., Smoot, G. F., Wilkinson, D. T., & Wright, E. L. 1992, *ApJ*, 397, 420
- Brown, J. C., Haverkorn, M., Gaensler, B. M., Taylor, A. R., Bizunok, N. S., McClure-Griffiths, N. M., Dickey, J. M., & Green, A. J. 2007, *ApJ*, 663, 258
- Burton, W. B., Deul, E. R., & Liszt, H. S. 1992, in *Saas-Fee Advanced Course of the Swiss Society for Astrophysics and Astronomy: The Galactic Interstellar Medium*, (New York: Springer-Verlag), 1
- Burton, W. B., Gordan, M. A., Bania, T. M., & Lockman, F. J. 1975, *ApJ*, 202, 30
- Buta, R., & Combes, F. 1996, *Fundamentals of Cosmic Physics*, 17, 95
- Carey, S. J., Noriega-Crespo, A., Mizuno, D. R., Shenoy, S., Paladini, R., Kraemer, K. E., Price, S. D., Flagey, N., Ryan, E., Ingalls, J. G., Kuchar, T. A., Gonçalves, D. P., Indebetouw, R., Billot, N., Marleau, F. R., Padgett, D. L., Rebull, L. M., Bressert, E., Ali, B., Molinari, S., Martin, P. G., Berriman, G. B., Boulanger, F., Latter, W. B., Miville-Deschenes, M. A., Shipman, R., & Testi, L. 2009, *PASP*, 121, 76
- Cazzolato, F., & Pineault, S. 2005, *AJ*, 129, 2731
- Churchwell, E. 2002, *ARAA*, 40, 27
- Churchwell, E., Babler, B. L., Meade, M. R., Whitney, B. A., Benjamin, R., Indebetouw, R., Cyganowski, C., Robitaille, T. P., Povich, M. S., Watson, C., & Bracker, S. 2009, *PASP*, 121, 213
- Churchwell, E., Povich, M. S., Allen, D., Taylor, M. G., Meade, M. R., Babler, B. L., Indebetouw, R., Watson, C., Whitney, B. A., Wolfire, M. G., Bania, T. M., Benjamin, R. A., Clemens, D. P., Cohen, M., Cyganowski, C. J., Jackson, J. M., Kobulnicky, H. A., Mathis, J. S., Mercer, E. P., Stolovy, S. R., Uzpen, B., Watson, D. F., & Wolff, M. J. 2006, *ApJ*, 649, 759
- Churchwell, E., Watson, D. F., Povich, M. S., Taylor, M. G., Babler, B. L., Meade, M. R., Benjamin, R. A., Indebetouw, R., & Whitney, B. A. 2007, *ApJ*, 670, 428
- Clark, J. S., Negueruela, I., Davies, B., Larionov, V. M., Ritchie, B. W., Figer, D. F., Messineo, M., Crowther, P. A., & Arkharov, A. A. 2009, *A&A*, 498, 109
- Cohen, M., Walker, R. G., Jayaraman, S., Barker, E., & Price, S. D. 2002, *AJ*, 121, 1180
- Cole, A. A., & Weinberg, M. D. 2002, *ApJ*, 574, L43
- Cragg, D. M., Johns, K. P., Godfrey, P. D., & Brown, R. D. 1992, *MNRAS*, 259, 203

- Cragg, D. M., Sobolev, A. M., & Godfrey, P. D. 2005, *MNRAS*, 360, 533
- Cyganowski, C. J., Brogan, C. L., Hunter, T. R., & Churchwell, E. 2009, *ApJ*, 702, 1615
- Cyganowski, C. J., Whitney, B. A., Holden, E., Braden, E., Brogan, C. L., Churchwell, E., Indebetouw, R., Watson, D. F., Babler, B. L., Benjamin, R., Gomez, M., Meade, M. R., Povich, M. S., Robitaille, T. P., & Watson, C. 2008, *AJ*, 136, 2391
- Dame, T. M., Hartmann, D., & Thaddeus, P. 2001, *ApJ*, 547, 792
- Dame, T. M., & Thaddeus, P. 2008, *ApJ*, 683, L143
- De Buizer, J. M. D., & Vacca, W. D. 2010, *AJ*, 140, 196
- Debattista, V. P., & Shen, J. 2007, *ApJ*, 654, L127
- Deharveng, L., Zavagno, A., Schuller, F., Caplan, J., Pomarès, M., & Breuck, C. D. 2009, *A&A*, 496, 177
- Devine, K. 2009, PhD Thesis, U of Wisconsin-Madison
- Diehl, R., Halloin, H., Kretschmer, K., Lichti, G. G., Schönfelder, V., Strong, A. W., von Kienlin, A., Wang, W., Jean, P., Knödlseeder, J., Roques, J.-P., Weidenspointner, G., Schanne, S., Hartmann, D. H., Winkler, C., & Wunderer, C. 2006, *Nature*, 439, 45
- Draine, B. T. 2003, *ARAA*, 41, 241
- Draine, B. T., Dale, D. A., Bendo, G., Gordon, K. D., Smith, J. D. T., Armus, L., Engelbracht, C. W., Helou, G., Kennicutt, R. C., Li, A., Roussel, H., Walter, F., Calzetti, D., Moustakas, J., Murphy, E. J., Rieke, G. H., Bot, C., Hollenbach, D. J., Sheth, K., & Teplitz, H. I. 2007, *ApJ*, 663, 866
- Draine, B. T., & Li, A. 2001, *ApJ*, 551, 807
- Drimmel, R. 2000, *A&A*, 358, L13
- Drimmel, R., Cabrera-Lavers, A., & López-Corredoira, M. 2003, *A&A*, 409, 205
- Drimmel, R., & Spergel, D. N. 2001, *ApJ*, 556, 181
- Durant, M., & van Kerkwijk, M. H. 2006, *ApJ*, 650, 1070
- Dwek, E., Arendt, R. G., Hauser, M. G., Kelsall, T., Lisse, C. M., Moseley, S. H., Silverberg, R. F., Sodroski, T. J., & Weiland, J. L. 1995, *ApJ*, 445, 716
- Elmegreen, B. G., & Lada, C. J. 1977, *ApJ*, 214, 725
- Englmaier, P., & Gerhard, O. 1999, *MNRAS*, 304, 512
- Erwin, P., Pohlen, M., & Beckman, J. E. 2008, *AJ*, 135, 20
- Everett, J. E., & Churchwell, E. 2010, *ApJ*, 713, 592
- Fich, M., Blitz, L., & Stark, A. A. 1989, *ApJ*, 342, 272
- Fouqué, P., Chevallerier, L., Cohen, M., Galliano, E., Loup, C., Alard, C., de Batz, B., Bertin, E., Borsenberger, J., Cioni, M. R., Copet, E., Dennefeld, M., Derriere, S., Deul, E., Duc, P.-A., Egret, D., Epchtein, N., Forveille, T., Garzón, F., Habing, H. J., Hron, J., Kimeswenger, S., Lacombe, F., Bertre, T. L., Mamon, G. A., Omont, A., Paturel, G., Pau, S., Persi, P., Robin, A. C., Rouan, D., Schultheis, M., Simon, G., Tiphène, D., Vauglin, I., & Wagner, S. J. 2000, *A&A Supp*, 141, 313
- Freeman, K. C. 1970, *ApJ*, 160, 811
- Freudenreich, H. T. 1998, *ApJ*, 492, 495
- Freudenreich, H. T., Berriman, G. B., Dwek, E., Hauser, M. G., Kelsall, T., Moseley, S. H., Silverberg, R. F., Sodroski, T. J., Toller, G. N., & Weiland, J. L. 1994, *ApJ*, 429, L69
- Fux, R. 1999, *A&A*, 345, 787
- Gao, J., Jiang, B. W., & Li, A. 2009, *ApJ*, 707, 89
- Georgelin, Y. M., & Georgelin, Y. P. 1976, *A&A*, 49, 57
- Gerhard, O. 2002, in *The Dynamics, Structure & History of Galaxies: A Workshop in Honour of Professor Ken Freeman*, eds. G.S. Da Costa & H. Jerjen, 273, 73
- Ghez, A. M., Salim, S., Weinberg, N. N., Lu, J. R., Do, T., Dunn, J. K., Matthews, K., Morris, M. R., Yelda, S., Becklin, E. E., Kremenek, T., Milosavljevic, M., & Naiman, J. 2008, *ApJ*, 689, 1044
- Gillessen, S., Eisenhauer, F., Fritz, T. K., Bartko, H., Dodds-Eden, K., Pfuhl, O., Ott, T., & Genzel, R. 2009, *ApJL*, 707, L114
- Girardi, L., Groenewegen, M. A. T., Hatziminaoglou, E., & da Costa, L. 2005, *A&A*, 436, 895
- Girardi, L., & Salaris, M. 2001, *MNRAS*, 323, 109
- Green, J. A., McClure-Griffiths, N. M., Caswell, J. L., Ellingsen, S. P., Fuller, G. A., Quinn, L., & Voronkov, M. A. 2009, *ApJL*, 696, L156
- Grocholski, A. J., & Sarajedini, A. 2002, *AJ*, 123, 1603
- Groenewegen, M. A. T. 2008, *A&A*, 488, 935
- Groenewegen, M. A. T., & Blommaert, J. A. D. L. 2005, *A&A*, 443, 143
- Gvaramadze, V. V., Kniazev, A. Y., & Fabrika, S. 2010, *MNRAS*, 405, 1047
- Habing, H. J. 1996, *Astronomy and Astrophysics Review*, 7, 97
- Hammersley, P. L., Garzón, F., Mahoney, T. J., López-Corredoira, M., & Torres, M. A. P. 2000, *MNRAS*, 317, L45
- Haslam, C. G. T., Salter, C. J., Stoffel, H., & Wilson, W. E. 1982, *Astronomy and Astrophysics Supplement Series*, 47, 1
- Heitsch, F., Whitney, B. A., Indebetouw, R., Meade, M. R., Babler, B. L., & Churchwell, E. 2007, *ApJ*, 656, 227
- Helfand, D. J., Becker, R. H., White, R. L., Fallon, A., & Tuttle, S. 2006, *AJ*, 131, 2525
- Hora, J. L., Latter, W. B., Allen, L. E., Marengo, M., Deutsch, L. K., & Pipher, J. L. 2004, *ApJSupp*, 154, 296
- Hora, J. L., Latter, W. B., Smith, H. A., & Marengo, M. 2006, *ApJ*, 652, 426

- Indebetouw, R., Mathis, J. S., Babler, B. L., Meade, M. R., Watson, C., Whitney, B. A., Wolff, M. J., Wolfire, M. G., Cohen, M., Bania, T. M., Benjamin, R. A., Clemens, D. P., Dickey, J. M., Jackson, J. M., Kobulnicky, H. A., Marston, A. P., Mercer, E. P., Stauffer, J. R., Stolovy, S. R., & Churchwell, E. 2005, *ApJ*, 619, 931
- Ishihara, D., Onaka, T., Kataza, H., Salama, A., Alfageme, C., Cassatella, A., Cox, N., García-Lario, P., Stephenson, C., Cohen, M., Fujishiro, N., Fujiwara, H., Hasegawa, S., Ita, Y., Kim, W., Matsuhara, H., Murakami, H., Müller, T. G., Nakagawa, T., Ohyama, Y., Oyabu, S., Pyo, J., Sakon, I., Shibai, H., Takita, S., Tanabé, T., Uemizu, K., Ueno, M., Usui, F., Wada, T., Watarai, H., Yamamura, I., & Yamauchi, C. 2010, *A&A*, 514, 1
- Jackson, J. M., Finn, S. C., Rathborne, J. M., Chambers, E. T., & Simon, R. 2008, *ApJ*, 680, 349
- Jackson, J. M., Rathborne, J. M., Shah, R. Y., Simon, R., Bania, T. M., Clemens, D. P., Chambers, E. T., Johnson, A. M., Dormody, M., Lavoie, R., & Heyer, M. H. 2006, *ApJSupp*, 163, 145
- Jiang, B. W., Gao, J., Omont, A., Schuller, F., & Simon, G. 2006, *A&A*, 446, 551
- Jiang, B. W., Omont, A., Ganesh, S., Simon, G., & Schuller, F. 2003, *A&A*, 400, 903
- Johnston, K. J., Gaume, R., Stolovy, S., Wilson, T. L., Walmsley, C. M., & Menten, K. M. 1992, *ApJ*, 385, 232
- Jurić, M., Ivezić, Ž., Brooks, A., Lupton, R. H., Schlegel, D., Finkbeiner, D., Padmanabhan, N., Bond, N., Sesar, B., Rockosi, C. M., Knapp, G. R., Gunn, J. E., Sumi, T., Schneider, D. P., Barentine, J. C., Brewington, H. J., Brinkmann, J., Fukugita, M., Harvanek, M., Kleinman, S. J., Krzesinski, J., Long, D., Neilsen, E. H., Nitta, A., Snedden, S. A., & York, D. G. 2008, *ApJ*, 673, 864
- Kent, S. M., Dame, T. M., & Fazio, G. 1991, *ApJ*, 378, 131
- Kerr, F. J., & Lynden-Bell, D. 1986, *MNRAS*, 221, 1023
- Kessler, M. F., Steinz, J. A., Anderegg, M. E., Clavel, J., Drechsel, G., Estaria, P., Faelker, J., Riedinger, J. R., Robson, A., Taylor, B. G., & de Ferrán, S. X. 1996, *A&A*, 315, L27
- Koenig, X. P., Allen, L. E., Gutermuth, R. A., Hora, J. L., Brunt, C. M., & Muzerolle, J. 2008, *ApJ*, 688, 1142
- Kormendy, J., & Kennicutt, R. C. 2004, *ARAA*, 42, 603
- Kurtz, S., Hofner, P., & Álvarez, C. V. 2004, *ApJSupp*, 155, 149
- Kwok, S. 2007, *Physics and Chemistry of the Interstellar Medium* (Sausalito, CA: University Science Books)
- Kwok, S., Zhang, Y., Koning, N., Huang, H.-H., & Churchwell, E. 2008, *ApJSupp*, 174, 426
- Launhardt, R., Zylka, R., & Mezger, P. G. 2002, *A&A*, 384, 112
- Liszt, H. S. 1985, in *The Milky Way Galaxy: Proceedings of the 106th Symposium* (Dordrecht: D. Reidel Publishing Co.), 283
- Lockman, F. J. 1980, *ApJ*, 241, 200
- López-Corredoira, M., Cabrera-Lavers, A., Garzón, F., & Hammersley, P. L. 2002, *A&A*, 394, 883
- López-Corredoira, M., Cabrera-Lavers, A., Mahoney, T. J., Hammersley, P. L., Garzón, F., & González-Fernández, C. 2007, *AJ*, 133, 154
- López-Corredoira, M., Hammersley, P. L., Garzón, F., Cabrera-Lavers, A., Castro-Rodríguez, N., Schultheis, M., & Mahoney, T. J. 2001, *A&A*, 373, 139
- Lucas, P. W., Hoare, M. G., Longmore, A., Schröder, A. C., Davis, C. J., Adamson, A., Bandyopadhyay, R. M., de Grijs, R., Smith, M., Gosling, A., Mitchison, S., Gáspár, A., Coe, M., Tamura, M., Parker, Q., Irwin, M., Hambly, N., Bryant, J., Collins, R. S., Cross, N., Evans, D. W., Gonzalez-Solares, E., Hodgkin, S., Lewis, J., Read, M., Riello, M., Sutorius, E. T. W., Lawrence, A., Drew, J. E., Dye, S., & Thompson, M. A. 2008, *MNRAS*, 391, 136
- Lutz, D., Feuchtgruber, H., Genzel, R., Kunze, D., Rigopoulou, D., Spoon, H. W. W., Wright, C. M., Egami, E., Katterloher, R., Sturm, E., Wierprecht, E., Sternberg, A., Moorwood, A. F. M., & de Graauw, T. 1996, *A&A*, 315, L269
- Marshall, D. J., Fux, R., Robin, A. C., & Reylé, C. 2008, *A&A*, 477, L21
- Marshall, D. J., Joncas, G., & Jones, A. P. 2009, *ApJ*, 706, 727
- Marshall, D. J., Robin, A. C., Reylé, C., Schultheis, M., & Picaud, S. 2006, *A&A*, 453, 635
- Martos, M., Hernandez, X., Yáñez, M., Moreno, E., & Pichardo, B. 2004, *MNRAS*, 350, L47
- Mauerhan, J. C., van Dyk, S. D., & Morris, P. W. 2009, *PASP*, 121, 591
- McClure-Griffiths, N. M., & Dickey, J. M. 2007, *ApJ*, 671, 427
- McClure-Griffiths, N. M., Dickey, J. M., Gaensler, B. M., & Green, A. J. 2004, *ApJ*, 607, L127
- Mellinger, A. 2009, *PASP*, 121, 1180
- Mercer, E. P., Clemens, D. P., Meade, M. R., Babler, B. L., Indebetouw, R., Whitney, B. A., Watson, C., Wolfire, M. G., Wolff, M. J., Bania, T. M., Benjamin, R. A., Cohen, M., Dickey, J. M., Jackson, J. M., Kobulnicky, H. A., Mathis, J. S., Stauffer, J. R., Stolovy, S. R., Uzpén, B., & Churchwell, E. B. 2005, *ApJ*, 635, 560
- Merrifield, M. R. 2004, in *Milky Way Surveys: The Structure and Evolution of our Galaxy*, *Proc. of*

- ASP Conference 317, eds. D. Clemens, R. Shah, & T. Brainerd (San Francisco: ASP), 289
- Mihalas, D., & Binney, J. 1981, *Galactic Astronomy* (San Francisco: W. H. Freeman & Co)
- Minier, V., Ellingsen, S. P., Norris, R. P., & Booth, R. S. 2003, *A&A*, 403, 1095
- Minniti, D., Lucas, P. W., Emerson, J. P., Saito, R. K., Hempel, M., Pietrukowicz, P., Ahumada, A. V., Alonso, M. V., Alonso-Garcia, J., Arias, J. I., Bandyopadhyay, R. M., Barbá, R. H., Barbu, B., Bedin, L. R., Bica, E., Borissova, J., Bronfman, L., Carraro, G., Catelan, M., Clariá, J. J., Cross, N., de Grijs, R., Dékány, I., Drew, J. E., Fariña, C., Feinstein, C., Lajús, E. F., Gamen, R. C., Geisler, D., Gieren, W., Goldman, B., Gonzalez, O. A., Gunthardt, G., Gurovich, S., Hambly, N. C., Irwin, M. J., Ivanov, V. D., Jordán, A., Kerins, E., Kinemuchi, K., Kurtev, R., López-Corredoira, M., Maccarone, T., Masetti, N., Merlo, D., Messineo, M., Mirabel, I. F., Monaco, L., Morelli, L., Padilla, N., Palma, T., Parisi, M. C., Pignata, G., Rejkuba, M., Roman-Lopes, A., Sale, S. E., Schreiber, M. R., Schröder, A. C., Smith, M., Sodr , L., Soto, M., Tamura, M., Tappert, C., Thompson, M. A., Toledo, I., Zoccali, M., & Pietrzynski, G. 2010, *New Astronomy*, 15, 433
- Misiriotis, A., Xilouris, E. M., Papamastorakis, J., Boumis, P., & Goudis, C. D. 2006, *A&A*, 459, 113
- Mizuno, D. R., Kraemer, K. E., Flagey, N., Billot, N., Shenoy, S., Paladini, R., Ryan, E., Noriega-Crespo, A., & Carey, S. J. 2010, *AJ*, 139, 1542
- Molinari, S., Swinyard, B., Bally, J., Barlow, M., Bernard, J.-P., Martin, P., Moore, T., Noriega-Crespo, A., Plume, R., Testi, L., Zavagno, A., Abergel, A., Ali, B., Andr , P., Baluteau, J.-P., Benedettini, M., Bern , O., Billot, N. P., Blommaert, J., Bontemps, S., Boulanger, F., Brand, J., Brunt, C., Burton, M., Campeggio, L., Carey, S., Caselli, P., Cesaroni, R., Cernicharo, J., Chakrabarti, S., Chrysostomou, A., Codella, C., Cohen, M., Compiegne, M., Davis, C. J., de Bernardis, P., de Gasperis, G., Francesco, J. D., di Giorgio, A. M., Elia, D., Faustini, F., Fischera, J. F., Fukui, Y., Fuller, G. A., Ganga, K., Garcia-Lario, P., Giard, M., Giardino, G., Glenn, J., Goldsmith, P., Griffin, M., Hoare, M., Huang, M., Jiang, B., Joblin, C., Joncas, G., Juvela, M., Kirk, J., Lagache, G., Li, J. Z., Lim, T. L., Lord, S. D., Lucas, P. W., Maiolo, B., Marengo, M., Marshall, D., Masi, S., Massi, F., Matsuura, M., Meny, C., Minier, V., Miville-Desch nes, M.-A., Montier, L., Motte, F., M ller, T. G., Natoli, P., Neves, J., Olmi, L., Paladini, R., Paradis, D., Pestalozzi, M., Pezzuto, S., Piacentini, F., Pomar s, M., Popescu, C. C., Reach, W. T., Richer, J., Ristorcelli, I., Roy, A., Royer, P., Russeil, D., Saraceno, P., Sauvage, M., Schilke, P., Schneider-Bontemps, N., Schuller, F., Schultz, B., Shepherd, D. S., Sibthorpe, B., Smith, H. A., Smith, M. D., Spinoglio, L., Stamatellos, D., Strafella, F., Stringfellow, G., Sturm, E., Taylor, R., Thompson, M. A., Tuffs, R. J., Umama, G., Valenziano, L., Vavrek, R., Viti, S., Waelkens, C., Ward-Thompson, D., White, G., Wyrowski, F., Yorke, H. W., & Zhang, Q. 2010, *PASP*, 122, 314
- Momany, Y., Zaggia, S., Gilmore, G., Piotto, G., Carraro, G., Bedin, L. R., & de Angeli, F. 2006, *A&A*, 451, 515
- Morgan, W. W., Whitford, A. E., & Code, A. D. 1953, *ApJ*, 118, 318
- Murray, N., & Rahman, M. 2010, *ApJ*, 709, 424
- Neugebauer, G., Habing, H. J., van Duinen, R., Aumann, H. H., Baud, B., Beichman, C. A., Beintema, D. A., Boggess, N., Clegg, P. E., de Jong, T., Emerson, J. P., Gautier, T. N., Gillett, F. C., Harris, S., Hauser, M. G., Houck, J. R., Jennings, R. E., Low, F. J., Marsden, P. L., Miley, G., Olmon, F. M., Pottasch, S. R., Raimond, E., Rowan-Robinson, M., Soifer, B. T., Walker, R. G., Wesselius, P. R., & Young, E. 1984, *ApJ*, 278, L1
- Nishiyama, S., Nagata, T., Baba, D., Haba, Y., Kadowaki, R., Kato, D., Kurita, M., Nagashima, C., Nagayama, T., Murai, Y., Nakajima, Y., Tamura, M., Nakaya, H., Sugitani, K., Naoi, T., Matsunaga, N., Tanab , T., Kusakabe, N., & Sato, S. 2005, *ApJ*, 621, L105
- Nishiyama, S., Nagata, T., Sato, S., Kato, D., Nagayama, T., Kusakabe, N., Matsunaga, N., Naoi, T., Sugitani, K., & Tamura, M. 2006, *ApJ*, 647, 1093
- Ohta, K., Hamabe, M., & Wakamatsu, K.-I. 1990, *ApJ*, 357, 71
- Omont, A., Gilmore, G. F., Alard, C., Aracil, B., August, T., Baliyan, K., Beaulieu, S., B gon, S., Bertou, X., Blommaert, J. A. D. L., Borsenberger, J., Burgdorf, M., Caillaud, B., Cesarsky, C., Chitre, A., Copet, E., de Batz, B., Egan, M. P., Egret, D., Epchtein, N., Felli, M., Fouqu , P., Ganesh, S., Genzel, R., Glass, I. S., Gredel, R., Groenewegen, M. A. T., Guglielmo, F., Habing, H. J., Hennebelle, P., Jiang, B., Joshi, U. C., Kimeswenger, S., Messineo, M., Miville-Desch nes, M. A., Moneti, A., Morris, M., Ojha, D. K., Ortiz, R., Ott, S., Parthasarathy, M., P rault, M., Price, S. D., Robin, A. C., Schultheis, M., Schuller, F., Simon, G., Soive, A., Testi, L., Teyssier, D., Tiph ne, D., Unavane, M., van Loon, J. T., & Wyse, R. 2003, *A&A*, 403, 975
- Oort, J. H. 1977, *ARAA*, 15, 295
- Paczynski, B. 1997, in *Variables Stars and the Astrophysical Returns of the Microlensing Surveys*,

- eds. R. Ferlet, J.-P. Maillard, & B. Raban (Gif-sur-Yvette, France: Editions Frontieres), 357
- Peretto, N. & Fuller, G. A. 2009, *A&A*, 505, 405
- Phillips, J. P., & Ramos-Larios, G. 2008a, *MNRAS*, 386, 995
- Phillips, J. P., & Ramos-Larios, G. 2008b, *MNRAS*, 383, 1029
- Pidopryhora, Y., Lockman, F. J., & Shields, J. C. 2007, *ApJ*, 656, 928
- Plambeck, R. L., & Menten, K. M. 1990, *ApJ*, 364, 555
- Povich, M. S., Churchwell, E., Bieging, J. H., Kang, M., Whitney, B. A., Brogan, C. L., Kulesa, C. A., Cohen, M., Babler, B. L., Indebetouw, R., Meade, M. R., & Robitaille, T. P. 2009, *ApJ*, 696, 1278
- Povich, M. S., Stone, J. M., Churchwell, E., Zweibel, E. G., Wolfire, M. G., Babler, B. L., Indebetouw, R., Meade, M. R., & Whitney, B. A. 2007, *ApJ*, 660, 346
- Price, S. D., Egan, M. P., Carey, S. J., Mizuno, D. R., & Kuchar, T. A. 2001, *AJ*, 121, 2819
- Purcell, C. R., Hoare, M. G., & Diamond, P. 2008, in *Massive Star Formation: Observations Confront Theory: ASP Conference Series 387*, eds. H. Beuther, H. Linz, & T. Henning (San Francisco: ASP), 389
- Ragan, S. E., Bergin, E. A., & Gutermuth, R. A. 2009, *ApJ*, 698, 324
- Reach, W. T., Rho, J., Tappe, A., Pannuti, T. G., Brogan, C. L., Churchwell, E. B., Meade, M. R., Babler, B., Indebetouw, R., & Whitney, B. A. 2006, *AJ*, 131, 1479
- Reid, M. J. 1993, *ARAA*, 31, 345
- Reid, M. J., Menten, K. M., Zheng, X. W., Brunthaler, A., & Xu, Y. 2009, *ApJ*, 705, 1548
- Reipurth, B. 2008, *Handbook of Star Forming Regions, Vol. 1: The Northern Sky* (San Francisco: ASP Monograph Publications)
- Reyl e, C., Marshall, D. J., Robin, A. C., & Schultheis , M. 2009, *A&A*, 495, 819
- Rhoads, J. E. 1998, *AJ*, 115, 472
- Rieke, G. H., Blaylock, M., Decin, L., Engelbracht, C., Ogle, P., Avrett, E., Carpenter, J., Cutri, R. M. et al. 1995, *AJ*, 135, 2245
- Rix, H.-W., & Zaritsky, D. 1995, *ApJ*, 447, 82
- Robin, A. C. 2009, *A&A*, 500, 165
- Robin, A. C., Creze, M., & Mohan, V. 1992a, *ApJ*, 400, L25
- Robin, A. C., Creze, M., & Mohan, V. 1992b, *A&A*, 265, 32
- Robin, A. C., Reyle, C., Derriere, S. & Picaud, S. 2003, *A&A*, 409, 523
- Robitaille, T. P., Cohen, M., Whitney, B. A., Meade, M., Babler, B., Indebetouw, R., & Churchwell, E. 2007a, *AJ*, 134, 2099
- Robitaille, T. P., Whitney, B. A., Indebetouw, R., & Wood, K. 2007b, *ApJSupp*, 169, 328
- Robitaille, T. P., Meade, M. R., Babler, B. L., Whitney, B. A., Johnston, K. G., Indebetouw, R., Cohen, M., Povich, M. S., Sewilo, M., Benjamin, R. A., & Churchwell, E. 2008, *AJ*, 136, 2413
- Robitaille, T. P., & Whitney, B. A. 2010, *ApJL*, 710, L11
- Robitaille, T. P., Churchwell, E., Benjamin, R. A., Whitney, B. A., Wood, K., Babler, B. L., & Meade, M. R. 2012, *A&A*, in press
- Rodriguez-Fernandez, N. J., & Combes, F. 2008, *A&A*, 489, 115
- Ruphy, S., Robin, A. C., Epchtein, N., Copet, E., Bertin, E., Fouque, P., & Guglielmo, F. 1996, *A&A*, 313, L21
- Sackett, P. D. 1997, *ApJ*, 483, 103
- Salpeter, E. E. 1955, *ApJ*, 121, 161
- Sawada, T., Hasegawa, T., Handa, T., & Cohen, R. J. 2004, *MNRAS*, 349, 1167
- Schweizer, F. 1976, *ApJ Supp*, 31, 313
- Scoville, N. Z., & Solomon, P. M. 1975, *ApJ*, 199, L105
- Seigar, M. S., & James, P. A. 1998, *MNRAS*, 299, 685
- Sevenster, M. N. 1999, *MNRAS*, 310, 629
- Shu, F. H., Milione, V., & Roberts, W. W. 1973, *ApJ*, 183, 819
- Simon, R., Jackson, J. M., Rathborne, J. M., & Chambers, E. T. 2006, *ApJ*, 639, 227
- Skrutskie, M. F., Cutri, R. M., Stiening, R., Weinberg, M. D., Schneider, S., Carpenter, J. M., Beichman, C., Capps, R., Chester, T., Elias, J., Huchra, J., Liebert, J., Lonsdale, C., Monet, D. G., Price, S., Seitzer, P., Jarrett, T., Kirkpatrick, J. D., Gizis, J. E., Howard, E., Evans, T., Fowler, J., Fullmer, L., Hurt, R., Light, R., Kopan, E. L., Marsh, K. A., McCallon, H. L., Tam, R., Dyk, S. V., & Wheelock, S. 2006, *AJ*, 131, 1163
- Smith, L. F., Biermann, P., & Mezger, P. G. 1978, *A&A*, 66, 65
- Srinivasan, S., Meixner, M., Leitherer, C., Vijn, U., Volk, K., Blum, R. D., Babler, B. L., Block, M., Bracker, S., Cohen, M., Engelbracht, C. W., For, B.-Q., Gordon, K. D., Harris, J., Hora, J. L., Indebetouw, R., Markwick-Kemper, F., Meade, M., Misselt, K. A., Sewilo, M., & Whitney, B. 2009, *AJ*, 137, 4810
- Stead, J. J., & Hoare, M. G. 2010, *MNRAS*, 407, 923
- Su, K. Y. L., Chu, Y.-H., Rieke, G. H., Huggins, P. J., Gruendl, R., Napiwotzki, R., Rauch, T., Latter, W. B., & Volk, K. 2007, *ApJ*, 657, L41
- Su, K. Y. L., Kelly, D. M., Latter, W. B., Misselt, K. A., Frank, A., Volk, K., Engelbracht, C. W., Gordon, K. D., Hines, D. C., Morrison, J. E., Muzerolle, J., Rieke, G. H., Stansberry, J. A., & Young, E. 2004, *ApJSupp*, 154, 302
- Tielens, A. G. G. M. 2005, *The Physics and Chemistry of the Interstellar Medium*, (Cambridge, UK: Cambridge University Press)

- Ueta, T. 2006, *ApJ*, 650, 228
- van der Kruit, P. C., & Searle, L. 1981, *A&A*, 95, 105
- van Loon, J. T., Gilmore, G. F., Omont, A., Blommaert, J. A. D. L., Glass, I. S., Messineo, M., Schuller, F., Schultheis, M., Yamamura, I., & Zhao, H. S. 2003, *MNRAS*, 338, 857
- van Woerden, H., Rougoor, G. W., & Oort, J. H. 1957, *Comptes Rendus l'Academie des Sciences*, 244, 1691
- Vanhollebeke, E., Groenewegen, M. A. T., & Girardi, L. 2009, *A&A*, 498, 95
- Vig, S., Ghosh, S. K., & Ojha, D. K. 2005, *A&A*, 436, 867
- Wachter, S., Mauerhan, J. C., van Dyk, S. D., Hoard, D. W., Kafka, S., & Morris, P. W. 2010, *AJ*, 139, 2330
- Wainscoat, R. J., Cohen, M., Volk, K., Walker, H. J., & Schwartz, D. E. 1992, *ApJSupp*, 83, 111
- Ward-Thompson, D. 1994, in *Clouds; cores and low mass stars: ASP Conference Series 65*, eds. D. Clemens & R. Barvainis (San Francisco: ASP), 207
- Watson, C., Povich, M. S., Churchwell, E. B., Babler, B. L., Chunev, G., Hoare, M., Indebetouw, R., Meade, M. R., Robitaille, T. P., & Whitney, B. A. 2008, *ApJ*, 681, 1341
- Weinberg, M. D. 1992, *ApJ*, 384, 81
- Weingartner, J. C., & Draine, B. T. 2001, *ApJ*, 548, 296
- Westerhout, G. 1957, *Bulletin of the Astronomical Institutes of the Netherlands*, 13, 201
- Whitney, B. 2009, *BAAS*, 41, 715
- Whitney, B. A., Indebetouw, R., Bjorkman, J. E., & Wood, K. 2004, *ApJ*, 617, 1177
- Whittet, D. C. B. 2003, *Dust in the galactic environment* (Bristol: IOP Publishing)
- Williams, J. P., de Geus, E. J., & Blitz, L. 1994, *ApJ*, 428, 693
- Wozniak, P. R., Udalski, A., Szymanski, M., Kubiak, M., Pietrzynski, G., Soszynski, I., & Zebrun, K. 2002, *Acta Astronomica*, 52, 129
- Wright, E. L., Eisenhardt, P. R. M., Mainer, A., Ressler, M. E., Cutri, R. M., Jarrett, T., Kirkpatrick, J. D., et al. 2010, *AJ*, 140, 1868
- Yanny, B., Rockosi, C., Newberg, H. J., Knapp, G. R., Adelman-McCarthy, J. K., Alcorn, B., Allam, S., Prieto, C. A., An, D., Anderson, K. S. J., Anderson, S., Bailer-Jones, C. A. L., Bastian, S., Beers, T. C., Bell, E., Belokurov, V., Bizyaev, D., Blythe, N., Bochanski, J. J., Boroski, W. N., Brinchmann, J., Brinkmann, J., Brewington, H., Carey, L., Cudworth, K. M., Evans, M., Evans, N. W., Gates, E., Gänsicke, B. T., Gillespie, B., Gilmore, G., Gomez-Moran, A. N., Grebel, E. K., Greenwell, J., Gunn, J. E., Jordan, C., Jordan, W., Harding, P., Harris, H., Hendry, J. S., Holder, D., Ivans, I. I., Ivezić, Ž., Jester, S., Johnson, J. A., Kent, S. M., Kleinman, S., Kniazev, A., Krzesinski, J., Kron, R., Kuropatkin, N., Lebedeva, S., Lee, Y. S., Leger, R. F., Lépine, S., Levine, S., Lin, H., Long, D. C., Loomis, C., Lupton, R., Malanushenko, O., Malanushenko, V., Margon, B., Martinez-Delgado, D., McGehee, P., Monet, D., Morrison, H. L., Munn, J. A., Neilsen, E. H., Nitta, A., Norris, J. E., Oravetz, D., Owen, R., Padmanabhan, N., Pan, K., Peterson, R. S., Pier, J. R., Platson, J., Fiorentin, P. R., Richards, G. T., Rix, H.-W., Schlegel, D. J., Schneider, D. P., Schreiber, M. R., Schwobe, A., Sibley, V., Simmons, A., Snedden, S. A., Smith, J. A., Stark, L., Stauffer, F., Steinmetz, M., Stoughton, C., Rao, M. S., Szalay, A., Szkody, P., Thakar, A. R., Thirupathi, S., Tucker, D., Uomoto, A., Berk, D. V., Vidrih, S., Wadadekar, Y., Watters, S., Wilhelm, R., Wyse, R. F. G., Yarger, J., & Zucker, D. 2009, *AJ*, 137, 4377
- Zasowski, G., Majewski, S. R., Indebetouw, R., Meade, M. R., Nidever, D. L., Patterson, R. J., Babler, B., Skrutskie, M. F., Watson, C., Whitney, B. A., & Churchwell, E. 2009, *ApJ*, 707, 510
- Zoccali, M. 2010, in *Chemical Abundances in the Universe: Connecting First Stars to Planets*, IAU Symposium 265, 271
- Zwicky, F. 1955, *PASP*, 67, 232

10 Interstellar PAHs and Dust

A. G. G. M. Tielens

Leiden Observatory, Leiden University, Leiden, The Netherlands

1	<i>Introduction</i>	500
2	<i>Observations of Interstellar PAHs and Dust</i>	501
2.1	Interstellar Extinction	501
2.2	IR Emission	504
2.3	Stardust	505
2.4	Interstellar Depletion	506
3	<i>Characteristics of Interstellar Dust</i>	507
3.1	Extinction by Grains	507
3.2	The Sizes of Interstellar Grains	509
3.3	Composition	513
4	<i>Physics of Interstellar PAHs and Dust</i>	516
4.1	Temperature and Infrared Emission	516
4.1.1	Radiative Temperature	516
4.1.2	Temperature Fluctuations	517
4.1.3	Temperature Distribution Function and PAH Spectrum	519
4.1.4	IR Emission Models	522
4.2	Charge	523
4.2.1	Collisional Rates	524
4.2.2	Photoelectric Rates	525
4.2.3	The Charge Distribution Function	526
5	<i>The Life Cycle of Interstellar PAHs and Dust</i>	527
5.1	Sources of Interstellar Dust	528
5.2	Formation in Stellar Ejecta	529
5.3	Processing by Interstellar Shocks	532
5.4	Depletions and the Life Cycle of Dust	535
6	<i>The Role of PAHs and Dust in the ISM</i>	538
6.1	Photoelectric Heating of Interstellar Gas	538
6.2	Interstellar Molecules	539
6.2.1	H ₂ Formation	539
6.2.2	Interstellar Ices	541
6.2.3	PAHs and Interstellar Molecules	543
7	<i>Conclusions and Outlook</i>	545
	<i>Acknowledgements</i>	546
	<i>References</i>	547

Abstract: Interstellar dust and large polycyclic aromatic hydrocarbon (PAHs) molecules are important components of the Interstellar Medium of galaxies where, among other things, they regulate the opacity, influence the heating and cooling of neutral atomic and molecular gas, and provide active surfaces for chemistry. Through this interaction with gas, photons, and energetic ions, dust and polycyclic aromatic hydrocarbon molecules influence key processes in the evolution of the interstellar medium and in turn are modified in their physical and chemical properties. This complex feedback drives the evolution of galaxies and its observational characteristics. In this chapter, our understanding of interstellar dust and large polycyclic aromatic hydrocarbon molecules is described. Besides observations and their analysis, this chapter describes the physical processes involved, the life cycle of interstellar dust, and some aspects of the role of interstellar dust and PAHs in the evolution of the interstellar medium.

1 Introduction

The presence of small dust grains in the interstellar medium of galaxies near and far is very apparent through extinction of stellar and nebular photons, through scattered light, through optical and infrared polarization, and through infrared emission. Polycyclic aromatic hydrocarbon (PAHs) molecules, the extension of the interstellar grain size distribution into the molecular domain, are equally prominent in images and spectra of galaxies at mid-infrared wavelengths. Moreover, many of the key processes that drive the evolution of galaxies – including star and planet formation as well as the accretion onto central black holes – occur deeply inside dust-enshrouded regions. Hence, understanding the characteristics of interstellar dust and PAHs is of key importance for our understanding of the evolution of the Universe.

The life cycle of the interstellar medium starts with the injection of material – much of it in the form of molecules or dust – by stars in the later stages of their life, the subsequent processing of this material in the interstellar medium by the prevalent ultraviolet radiation fields, energetic particles, and strong shocks, and ends with the incorporation of this material into newly formed stars and their budding planetary systems. During this evolution, the dust and molecules injected into the interstellar medium are modified or even completely transformed. Conversely, the dust and molecules injected into the ISM have a profound influence on their environment through a variety of processes and this complex feedback shapes the evolution of galaxies. Moreover, the effects of processes taking place in the stellar ejecta and in the interstellar medium will be inherited by newly forming planetary systems.

Interstellar dust and PAH molecules drive a number of key processes in the interstellar medium. Dust is the dominant opacity source from far ultraviolet to sub-millimeter wavelengths and controls therefore the spectral energy distribution over much of the observable range. Over this wavelength range, gas does not couple efficiently to the energy available in the photon field and the photoelectric effect on small dust grains and large molecules is thought to be the dominant heating source of neutral atomic or molecular gas, such as diffuse HI clouds and photodissociation regions associated with massive stars. Dust is also a major reservoir of the elements and typically ~90% of the abundance of most elements (except H, N, O, and the noble gases) are in solid form. In some cases (e.g., Ti, Ca, Fe), this depletion reaches 99–99.99% of the available atoms. The level of depletion has of course also indirect effects on, for example, the cooling of the gas and the resulting emission spectrum. Dust grains are also key to the formation and survival of molecules. First, dust has an indirect influence on molecular abundances

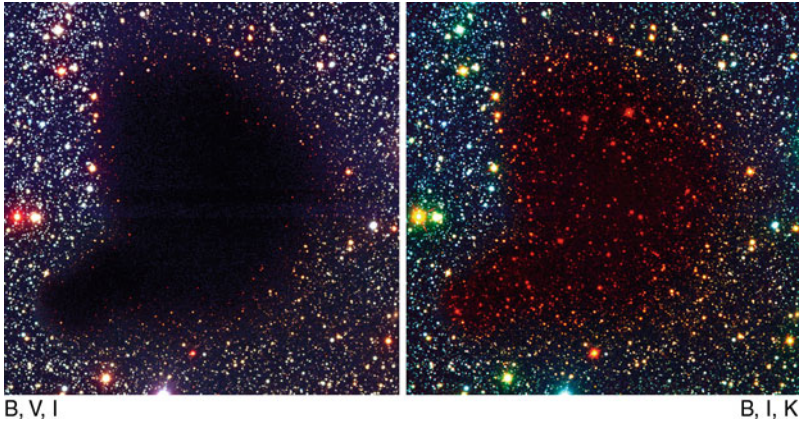
through its shielding effects on the penetrating far ultraviolet photons that are an important destruction agent of molecules. Second, the most abundant molecule, H_2 , is very inefficiently formed through direct gas phase processes and is therefore generally thought to be formed primarily on grain surfaces. The presence of copious amounts of H_2 is a prerequisite for efficient molecule formation in most gas phase chemical routes and so, from this point of view, grains play an important role in interstellar chemistry. Third, grain surfaces act as a “waterhole” where gaseous species meet and mate. Observations show that ices – consisting of simple molecules – are ubiquitous inside dense clouds and these ices reflect this active grain surface chemistry. This ice formation process is very efficient, removing most of the species except H_2 and He, from the gas phase in dense, prestellar cores. Again, this depletion affects the cooling ability of the gas. Besides their influence on molecular abundances, (charged) dust grains and large molecules also provide an efficient recombination route for gaseous ions and hence influence gas phase (ion-molecule) chemistry that way as well. In addition, ionization couples gas to the omnipresent magnetic field. These effects on the ionization and energy balance of interstellar gas play an important role in the star formation process. Furthermore, the dust-forming elements are observed to be overabundant in cosmic rays and, hence, dust may well be involved in the origin of cosmic rays through Fermi acceleration and sputtering of large dust grains across supernova shock fronts. Finally, dust grains are the fundamental building blocks of planetary systems. Indeed, studies of meteorites, interplanetary dust particles, and cometary dust particles have revealed the presence of genuine stardust grains which have survived the rigors of the ISM and the star and planet formation process and were captured in these Solar system objects without losing their (full) identity.

In this chapter, the characteristics of interstellar dust and PAH molecules are reviewed. **Section 2** focuses on the observations of interstellar dust, including extinction and IR emission. Other studies of interstellar dust are also briefly reviewed. These observations are interpreted in terms of sizes and composition in **Sect. 3**. The physics of interstellar dust and PAHs is often connected to their temperature and charge. Processes involved in the energy and ionization balance of these species are described in **Sect. 4**. The life cycle of interstellar dust is discussed in **Sect. 5** while several aspects of the role of interstellar dust and PAHs in the interstellar medium are described in **Sect. 6**. The references interspersed in the text had to be limited and do not fully do justice to this rich area of research. The reader may want to consult the monographs (Krügel 2002; Whitter 2003), that present in-depth discussions of observational and theoretical aspects of interstellar dust. The reviews (Draine 2003; Tielens 2008), as well as the proceedings of various conferences on interstellar dust (Allamandola and Tielens 1989; Henning et al. 2009; Witt et al. 2004) will provide further entries into the field.

2 Observations of Interstellar PAHs and Dust

2.1 Interstellar Extinction

The presence of interstellar dust first manifested itself through extinction of starlight and dark clouds are obvious examples of this (**Fig. 10-1**) as noticed, e.g., by William Herschel as he exclaimed “Hier ist wahrhaftig ein Loch im Himmel” (there is truly a hole in the sky) to his sister, and companion in a life of astronomical discoveries, Caroline, after observing a dark cloud near Scorpius.



■ Fig. 10-1

A comparison of the dark globule, B68, in a color composite of visible (B & V) and near-infrared (I) on the *left* and a false-color composite based on a visible (V, here rendered as blue), a near-infrared (I, green) and an infrared (K, red) on the *right*. The effect of dust extinction is obvious: while at visual wavelengths, this globule is *black*, at IR wavelengths background stars shine through

Extinction, A , is related to the apparent magnitude of a star, m , through

$$m(\lambda) = M(\lambda) + 5 \log [d] + A_\lambda, \quad (10.1)$$

with M the absolute magnitude of the star and d the distance. Extinction is thus measured on a magnitude scale. In terms of optical depth, $\tau_\lambda = 0.92A_\lambda$. Quantitatively, the color excess, $E(\lambda - V) = A_\lambda - A_V$, can then be measured by comparing the light of a dust-obscured star with that of an unobscured star with the same spectral type and luminosity class. Throughout the visible, the color excess generally increases with wavelength and hence extinction is often called reddening. Assuming that the extinction goes to zero at infrared wavelength, the ratio of total to selective extinction, $R_V = A_V/E(B - V)$, can be determined. The extinction ratio, A_λ/A_V , which is more readily related to the physical properties of dust, can then be determined from the observations; namely,

$$\frac{A_\lambda}{A_V} = \frac{1}{R_V} \frac{E(\lambda - V)}{E(B - V)} + 1. \quad (10.2)$$

► *Figure 10-2* shows galactic extinction curves measured along “quiescent” sight lines in the local Solar neighborhood (~ 1 kpc). In general, these extinction curves are characterized by a rapidly rising extinction curve in the infrared and visual parts of the spectrum, scaling with $\lambda^{-1.7}$ and λ^{-1} , respectively. In the UV, the extinction curve behavior changes character with a knee around $\lambda \simeq 2 \mu\text{m}^{-1}$, a pronounced bump at $\lambda = 4.67 \mu\text{m}^{-1}$ ($2,175 \text{ \AA}$), and a steep rise toward the far-UV. The widely varying behavior of interstellar extinction curves can be described by one parameter for which often R_V is chosen (Cardelli et al. 1989). The value of this parameter depends on the environment. The extinction curve in the diffuse ISM is often represented by $R_V = 3.1$, while in regions of star formation R_V is typically much larger (4–6). The extinction curves in the Magellanic cloud galaxies are extreme examples of these variations (► *Fig. 10-2*). Some sight lines – toward quiescent regions – in the LMC show extinction curves which are indistinguishable from the galactic sight lines described above. However, most of the extinction curves measured toward the Large and Small Magellanic Clouds are significantly

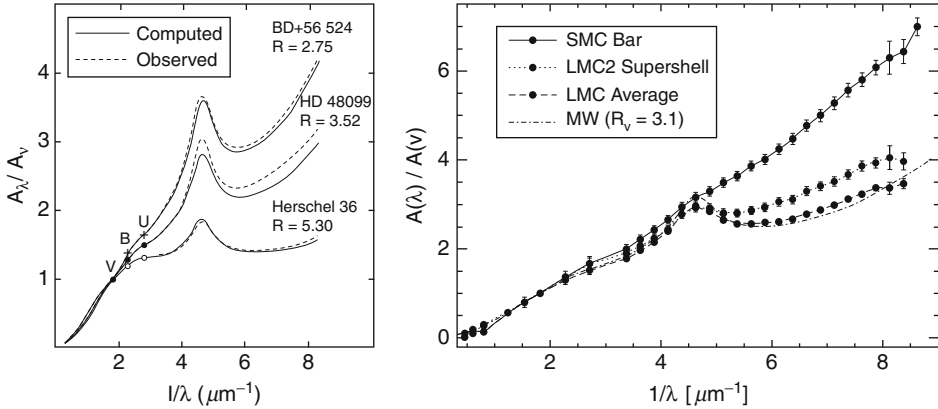


Fig. 10-2

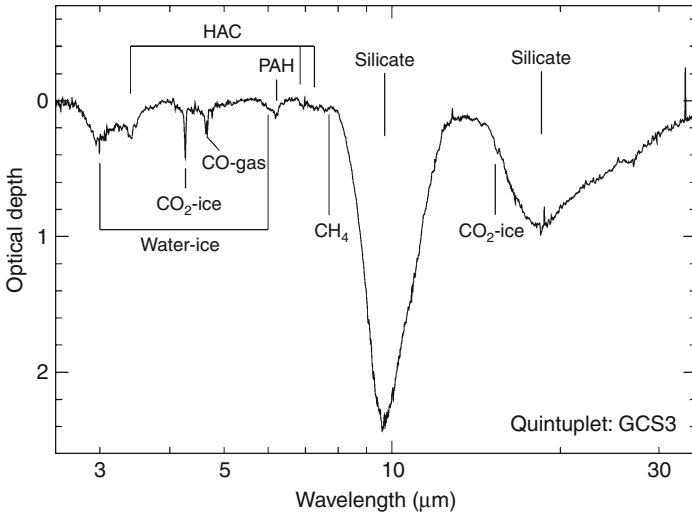
Left: Three observed galactic extinction curves as a function of inverse wavelength (e.g., frequency increases to the right). These curves illustrate the variation in the extinction law in the diffuse ISM of the Milky Way. The *solid lines* show commonly used parameterized fits (see text for details. Figure taken from Cardelli et al. 1989). **Right:** Extinction curves derived for the nearby galaxies, the Large and Small Magellanic clouds. For comparison, the “average” interstellar extinction curve in the Milky Way is also shown (Figure taken from Gordon et al. 2003)

different from the galactic ones (Fig. 10-2); characterized by weak or non-existent 2,175 Å bumps (Gordon et al. 2003). These extinction studies are biased toward active regions in the Clouds where shock processing may have affected the dust properties. Indeed, some low-density sight lines in the Milky Way – probing dust in the lower halo – are equally different from the average interstellar extinction curve measured in the local Solar neighborhood (Clayton et al. 2000) and likely probe the effects of recent shocks as well.

In the infrared, the extinction curve shows broad features due to vibrations in solid materials (Fig. 10-3) that can be used as “fingerprints” pinpointing the absorbing materials (Chiar et al. 2000). In the diffuse ISM, IR extinction is dominated by the broad and structureless 9.7 and 18 μm features – due to silicates – accompanied by much weaker features at 3.4, 6.8, and 7.2 μm – due to hydrogenated amorphous carbon. For sight lines through molecular clouds other absorption features appear (e.g., 3.07, 4.2, 4.67, 6.0, 6.85, and 15 μm) – due to simple molecules in an ice mixture – which show complex profile variations with location in the cloud (Boogert and Ehrenfreund 2004).

Besides through extinction of star light, the presence of dust is also apparent from scattered light in, for example, reflection nebulae. The scattering albedo is 0.5–0.6 throughout the visible, shows a drop at the position of the 2,175 Å bump, rises again to a value of ≈ 0.8 and then drops rapidly toward the far-UV. Dust also betrays its presence through polarization of starlight. This polarization peaks in the visible and drops rapidly toward both the infrared and the UV. The polarization behavior can also be well described by one parameter for which again R_V is often selected.

Lastly, the extinction is a measure for the dust column density and can be directly compared to the column density of atomic or molecular hydrogen. These are measured through their UV absorption bands (Lyman Alpha line for H and the Lyman-Werner bands for H₂). Extensive



■ Fig. 10-3

Extinction in the IR is characterized by a number of broad absorption bands due to solid state vibrations, here illustrated with the extinction profiles observed toward one of the quintuplet sources in the galactic center. This sight line traverses diffuse as well as molecular clouds. Indicated identifications refer to silicates and Hydrogenated Amorphous Carbon (HAC) in the diffuse ISM. In dense molecular clouds, simple molecules in ices are also present (Figure courtesy of Chiar et al. 2000)

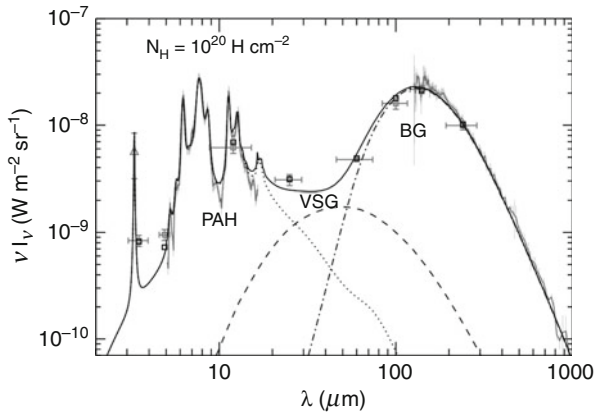
studies, dating back to the Copernicus satellite, have resulted in

$$\frac{N_H}{A_V} = 1.9 \times 10^{21} \text{ cm}^{-2} \text{ magn}^{-1}, \quad (10.3)$$

in the Solar neighborhood (Bohlin et al. 1978). On average, there is about two magnitudes of visual extinction per kpc. However, this extinction is very patchy (cf. Fig. 10-1) even in the diffuse ISM. Typically, there are about six “Spitzer-type” clouds with each $A_V \simeq 0.2$ and $N_H \simeq 4 \times 10^{20} \text{ cm}^{-2}$ and 0.8 larger clouds with $A_V \simeq 0.6 \text{ magn}$ and $N_H \simeq 1.2 \times 10^{21} \text{ cm}^{-2}$ per kpc (Bohlin et al. 1978).

2.2 IR Emission

Over the last three decades, satellite observations have opened up the IR spectral window for systematic studies and the IR emission (and absorption, cf. Fig. 10-3) of the ISM has become a major tool to study interstellar dust. The IR emission spectrum of the ISM (Fig. 10-4) is characterized by a broad emission peak at long wavelengths ($\simeq 150 \mu\text{m}$) characteristic for cold dust ($T \simeq 15 \text{ K}$). Near luminous stars, such as HII regions and reflection nebulae, the peak shifts to much shorter wavelength (up to $\sim 30 \mu\text{m}$; Peeters et al. 2002b). In addition to this bright far-IR peak, the IR emission spectrum of the ISM also shows bright mid-IR emission (Boulanger 2000). This mid-IR spectrum of the ISM is dominated by strong features at 3.3, 6.2, 7.7, 8.6, 11.2,



■ Fig. 10-4

Emission of dust in the diffuse high galactic latitude medium normalized for $N_H = 10^{20} \text{ H cm}^{-2}$. *Graysymbols and curves* display the emission spectrum observed with AROME balloon experiment, ISOCAM/CVF on board ISO, and DIRBE and FIRAS on board COBE. *Black lines* are the DUSTEM model output – adapted from the original model developed by Désert et al. (1990) – and *black squares* the modeled DIRBE points taking into account instrumental transmissions and color corrections. The fit – based upon the original model developed by Désert et al. (1990) – consists of (carbonaceous and silicate) Big Grains in radiative equilibrium responsible for the far infrared and submillimeter emission, (carbonaceous) Very Small Grains that fluctuate stochastically, responsible for the mid-IR continuum, and PAH molecules fluorescing in mid-IR emission bands (See Compiegne et al. 2010 for details and references)

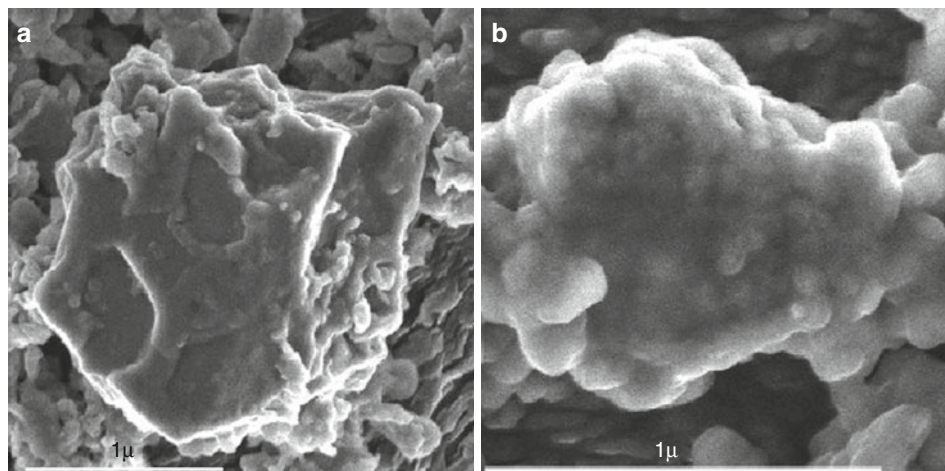
and $12.7 \mu\text{m}$ which are collectively known as the IR emission features¹ and a host of weaker features and underlying plateaus (Peeters et al. 2002a). These are perched on (quasi-)continuum emission which rises rapidly toward longer wavelengths.² The spectral characteristics of the IR emission features show subtle variations in profile, peak position, and relative strength within sources and from one source to another but overall the spectrum is remarkably similar; if you see one feature you see them all. In order to emit in the mid-IR, the carrier(s) of these features must be much hotter than the 15 K that characterizes the far-IR emission. The various components that can be recognized in these IR emission spectra – IR emission features, plateaus, and 25 and 60 μm “continuum” – are carried by independent but related species – PAH molecules, PAH clusters, very small grains, and small grains – each with their typical emission temperature ($\approx 600, 300, 150, 50 \text{ K}$).

2.3 Stardust

Genuine stardust grains have been isolated from carbonaceous meteorites. Analysis of these stardust grains has revealed an isotopic composition which is distinctly non-Solar and derives

¹In the older literature, these features are also referred to as the Unidentified InfraRed (UIR) bands.

²Presently features have been detected out to 20 μm but Herschel may extend this into the far-IR and even the sub-millimeter.



■ Fig. 10-5

Field-emission scanning electron microscope images of pristine presolar SiC grains from the Murchison meteorite (Bernatowicz et al. 2003): (a) exhibiting primary growth crystal faces and polygonal depressions; (b) coated with an apparently amorphous, possibly organic phase. Scale bars are 1 μm . Figure taken from Bernatowicz et al. (2003)

directly from the stellar birthsites of these dust grains (Anders and Zinner 1993). Apparently, these dust grains formed in stellar ejecta – enriched by the nucleosynthetic products of processes taking place in the deep interiors of these stars and then mixed to the surface – were injected into the ISM, processed by shocks and other energetic events, became part of a region of star formation that collapsed to form the Solar system, saw the hot gases swirl in the Solar nebula, experienced possibly the shocks and lightning processes rampant in this environment, were incorporated into the planetary body from which the meteorite was derived that crashed on Earth and then was analyzed in the laboratory; and through all this arduous and torturous history, these stardust grains never equilibrated fully with the gas and managed to preserve their stellar heritage. Detailed studies of these stardust grains have opened a new window on the dusty universe, providing new insights in the composition of dust and the processes that play a role in their formation (Bernatowicz et al. 2003). In addition, recent studies have started to explore the impact of interstellar processes on these grains (Henkel et al. 2007).

2.4 Interstellar Depletion

Elemental abundance studies in the interstellar medium have revealed that the abundance of almost all elements are less than in the Sun, in the Solar system (mainly meteorites), or in nearby stars.³ The “missing atoms” are thought to be locked up in dust and thus probe the dust composition. ▶ *Table 10-1* summarizes measured interstellar gas phase abundances for some of the major elements, relevant reference abundances, and derived elemental fractions locked up

³Solar, B-star, and young F-G star abundances differ by about 0.2 dex and moreover vary between different determinations (Jenkins 2009)

■ **Table 10-1**
Elemental depletions of the major dust-forming elements^a

Element	Solar abundances	Dust core abundances ^b	Dust abundances ^c
C	391	251	259
N	85	-8	9
O	490	170	170
Mg	35	12	32
Si	34	15	33
Fe	28	21	28

Ref. Savage and Sembach (1996)

^aPer million H-atoms

^bThe difference between Solar abundances and halo abundances (i.e., where only resilient dust cores are expected to survive)

^cThe difference between Solar abundances and abundances measured along the heavily depleted line of sight toward ζ Oph

in dust. The two entries refer to the “average” composition of resilient dust cores and of more “fragile” mantles (cf. ► Sect. 5.4). The total dust mass is then $7.6\text{--}9.2 \times 10^{-3}$ that of hydrogen. Because the reference elemental abundance scale is somewhat controversial and because these depletions are global values averaged over all dust compounds, it is difficult to translate these into a specific composition of the dust. The data is consistent, though, with a large contribution to the interstellar dust mass by oxides such as silicates and a more moderate carbon dust mass. However, whether, for example, iron is in the form of metallic grains or part of astronomical silicates, is unclear from depletion studies.

3 Characteristics of Interstellar Dust

3.1 Extinction by Grains

The optical depth due to dust is given by

$$\tau_d(\lambda) = \sum_i \int \int n_{d,i}(a) C_{\text{ext},i}(a, \lambda) da d\ell, \quad (10.4)$$

where the sum is over all dust types (materials), i , and the integrations are over the size distribution and the line of sight. Here, $n_{d,i}(a)$ is the density of dust grains of size a and $C_{\text{ext},i}(a, \lambda)$ the wavelength dependent extinction cross section. The latter can be written in terms of the extinction efficiency, Q_{ext} , normalized to the geometric cross section,

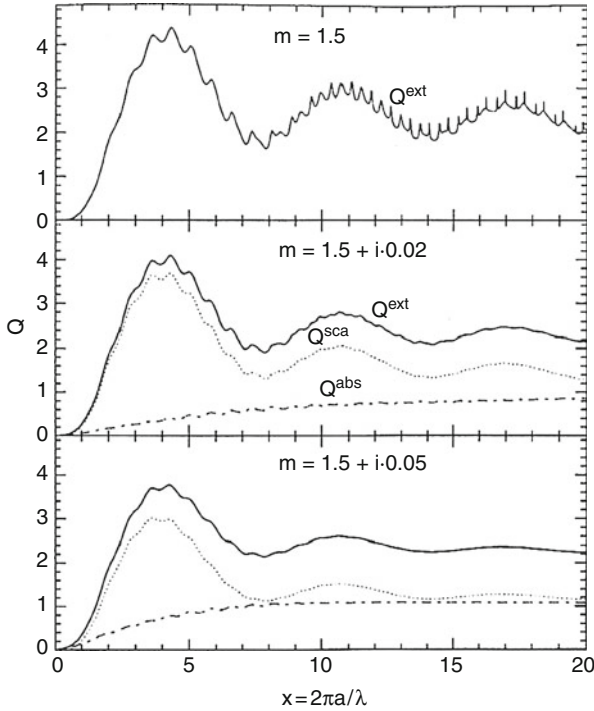
$$C_{\text{ext},i}(a, \lambda) = \pi a^2 Q_{\text{ext},i}(a, \lambda). \quad (10.5)$$

If the grain size distribution is approximated by a single size and a single type, the optical depth reduces to

$$\tau(\lambda) \simeq n_d \pi a^2 Q_{\text{ext}}(\lambda) L, \quad (10.6)$$

for a homogeneous density distribution along a sight line of length L .

The interaction of light with grains is described by the Maxwell equations with appropriate boundary conditions and the proper evaluation of grain cross sections is a topic in itself (Bohren



■ Fig. 10-6

Extinction and scattering efficiency, $Q (= C_{\text{ext}}/\pi a^2)$, for spherical grains as a function of the size parameter $2\pi a/\lambda$. The value for the complex index of refraction ($m = n + ik$ with n and k the real and imaginary part) is indicated in the panel (Figure taken from van de Hulst 1957)

and Huffman 1983; van de Hulst 1957). Dust extinction depends on many factors such as grain size, the optical properties of the constituent material(s), and the shape and morphology of a grain. Exact solutions have been derived for spherical grains and infinite cylinders. Approximate methods exist for grains of arbitrary shapes and composite grains. ● Figure 10-6 shows the result of such a calculation for a spherical grain with dielectric properties (e.g., silicate). For a dielectric grain, far from absorption resonances, the optical properties are quite constant and grain size is then the main parameter describing the extinction. In the Rayleigh limit (grain size, a , much less than the wavelength, λ ; e.g., $a \ll \lambda/2\pi$), extinction is dominated by absorption and the absorption cross section per unit volume is $\simeq (18\pi/\lambda) k/(n^2 + 2)^2$ (with $m = n + ik$ the complex index of refraction) and hence increases with λ^{-1} . The extinction cross section saturates and reaches a maximum when the phase lag, $\rho = (4\pi a/\lambda)(m-1)$, through the grain is $\simeq 4$. For larger grains (● Fig. 10-6), the dust cross section shows slow undulations related to interference fringes between the unperturbed wave and the forward scattered wave (e.g., essentially, Fresnel fringes). The rapid fringes are also due to interference but do not have such a “simple” interpretation. For large sizes, the absorption and scattering cross section both approach the geometric cross section: all the light falling on an object is absorbed while diffraction (scattering at small angles) at the grain edge “removes” also the equivalent of the grain cross section from the beam (Babinet’s principle, van de Hulst 1957).

Near a resonance, the optical constants of a material vary rapidly and these lead to strong variations in the wavelength dependence of extinction. For a dielectric material, resonances occur in the infrared due to vibrations of the atoms in the lattice and in the UV due to electronic transitions between the valence and the conduction band. The former are often represented by (collection of) Lorentz oscillators for which simple analytical expressions exist but it should be recognized that atoms in a lattice are not harmonic oscillators and this is only an approximation. The extinction (absorption) cross per unit volume for an arbitrary ellipsoid is

$$\frac{C_{\text{ext}}^j}{V} = \frac{2\pi}{\lambda} \frac{\varepsilon_2}{(L_j (\varepsilon_1 - 1) + 1)^2 + (L_j \varepsilon_2)^2}, \quad (10.7)$$

where the L_j are the depolarization factors that depend on shape ($L_j = 1/3$ for spheres) and ε is the (complex) dielectric constant of the material ($\varepsilon = m^2 = \varepsilon_1 + i\varepsilon_2$), which, for a Lorentz oscillator with resonance (circular) frequency, $\omega_o = 2\pi\nu_o$, is given by,

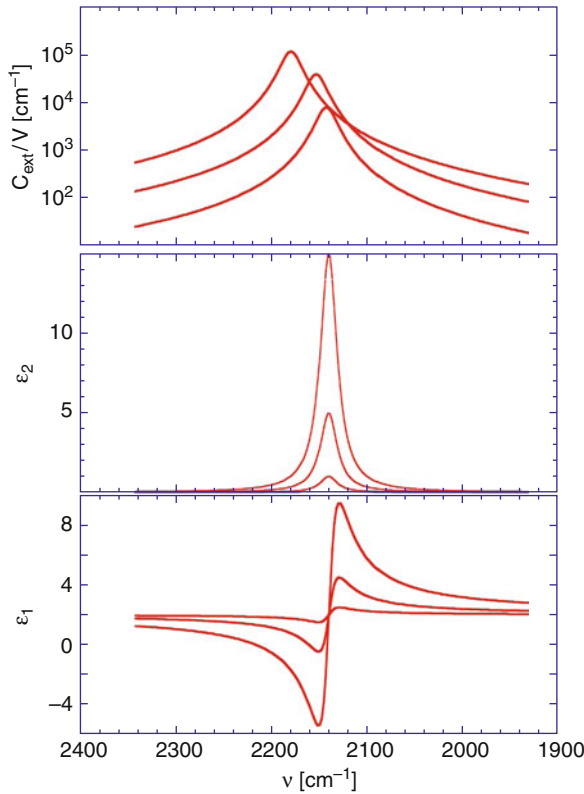
$$\varepsilon(\omega) = \varepsilon_\infty + \frac{S \omega_o^2}{\omega_o^2 - \omega^2 + i\omega\Gamma_o}, \quad (10.8)$$

with ε_∞ the high frequency dielectric constant, Γ_o the width, and S the strength of the transition.

The dielectric constants for a Lorentz oscillator with a transition frequency in the mid-IR are shown in **Fig. 10-7** for three different values of the intrinsic strength. The imaginary part shows the typical resonance of a damped oscillator while the real part shows a region of anomalous dispersion toward higher frequencies. As the strength of the oscillator increases, the variations in the dielectric constant become more pronounced. This figure also shows the extinction cross section per unit volume. For low strength, the cross section follows closely the variation in the imaginary part of the dielectric constant. However, when the strength increases, the real part can go negative and small particles can show resonances when $L_j(\varepsilon_1 - 1) + 1$ becomes zero. For a sphere, this resonance occurs when $\varepsilon_1 = -2$. As **Fig. 10-7** demonstrates, this can have a profound influence on the position and profile of strong absorption bands. In astrophysics, this plays, for example, a role in $\pi \rightarrow \pi^*$ transition in carbonaceous materials near 2,175 Å, the vibrational stretching modes in solid CO and CO₂, and the various lattice modes of crystalline silicates.

3.2 The Sizes of Interstellar Grains

The continued rise of the interstellar extinction curve (**Fig. 10-2**) through the infrared and visible and into the far-UV is a testimony to the presence of a broad size distribution, ranging from $a \sim 3,000 \text{ \AA}$ – responsible for the near-infrared extinction – to $a \sim 100 \text{ \AA}$ – responsible for the UV extinction. These extinction measurements can be readily translated into dust abundances once it is recognized that dust grains dominate extinction at wavelengths that match their size (e.g., $\lambda \sim 2\pi a$). Thus, the observed dust-to-gas ratio (cf. **10.3**) translates into a dust abundance of $1.5 \times 10^{-12} \text{ H-atom}^{-1}$ for the 1,000 Å grains responsible for the visual extinction (e.g., **10.6** with $2\pi a/\lambda \simeq 1$ and $Q \simeq 1$). From the far-UV extinction ($A_{\text{uv}}/A_V \simeq 4$; **Fig. 10-2**), a dust density is derived of $6 \times 10^{-10} \text{ H-atom}^{-1}$ for 100 Å grains. Similarly, ascribing the 2,175 Å bump to $\simeq 200 \text{ \AA}$ (graphite) grains and using $\Delta\tau_{2,175}/N_H \simeq 2 \times 10^{-21} (\text{H-atom})^{-1}$, the abundance is $1.5 \times 10^{-10} \text{ H-atom}^{-1}$. As this discussion shows, the interstellar grain size distribution is steeply rising toward small grains. The surface area is also dominated by the small grains, but the mass

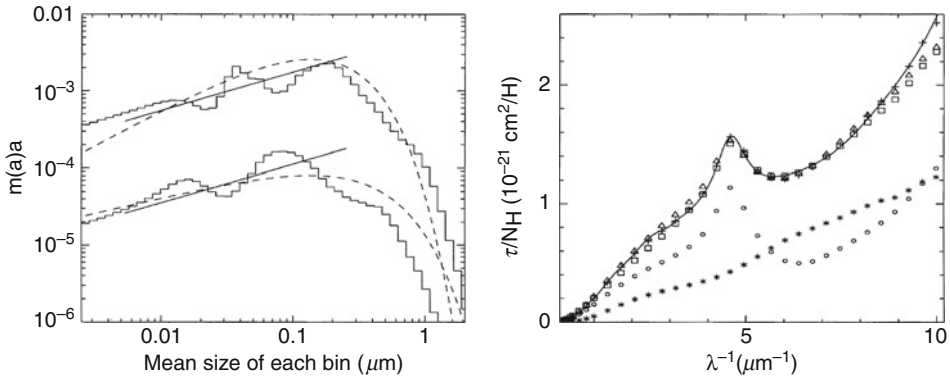


■ Fig. 10-7

The real (ϵ_1) and imaginary (ϵ_2) part of the dielectric constant of a Lorentz oscillator for different values of the strength ($S = 0.01, 0.05, 0.15$). *Top panel*: the extinction cross section per unit volume calculated for spheres using the dielectric constants given in the *bottom two panels*. Note that frequency increases to the *left*

is in large grains by a factor ~ 4 . With a typical specific density of astrophysically relevant grain materials of 2.5 g cm^{-3} , this analysis results in a dust-to-gas ratio of ~ 0.01 by mass.

Based upon the analysis of the extinction curve and with input from albedo studies, detailed models have been developed for the interstellar grain size distribution. These models assume specific grain materials – typically silicates and a form of carbonaceous material (graphite or amorphous carbon) – and the associated optical properties measured in the laboratory, as well as grain shape(s) – often spherical for simplicity or spheroidal if polarization is also modeled. With these assumptions, the extinction behavior of the dust can be calculated, using one of the various scattering codes around, and the observed extinction curve can then be inverted to arrive at the interstellar grain size distribution. None of these factors is well known and, hence, there is considerable ambiguity in the results, in a way reflecting personal taste of the theoreticians involved. All of these models agree in the global range of sizes present outlined above but not in the specifics. The most widely used model is the Mathis–Rumpl–Nordsieck model (or MRN model), named after the scientists who first derived this particular grain size distribution



■ Fig. 10-8

The interstellar grain size distribution – plotted as mass fractions – derived from different types of fits to the observed extinction curve (*left panel*). The *top histogram* and *dashed line* shows silicates; the *bottom histogram* and *dashed line* shows graphite, displaced downward by factor 10. The *solid lines* show for comparison the simple MRN power law size distributions. The calculated extinction curve is compared to the observations in the *right panel*. The contributions due to silicates (*) and graphite (o) are shown separately (Figure taken from Kim et al. 1994)

(Mathis et al. 1977). For all practical purposes, this model is as good as any other and it has the advantage of simplicity. This model consists of spherical graphite and silicate grains with a powerlaw size distribution with an exponent of -3.5 in the range $50\text{--}2,500\text{ \AA}$ given by

$$n_i(a) da = A_i n_H a^{-3.5} da, \quad (10.9)$$

with a the grain size, and where the constants A_i for silicate and graphite are given by $A_{\text{sil}} = 7.8 \times 10^{-26}$ and $A_{\text{gra}} = 6.9 \times 10^{-26} \text{ cm}^{2.5} (\text{H-atom})^{-1}$, respectively. Figure 10-8 shows a more involved size distribution derived from a least-square-fit to the observed interstellar extinction curve. Alternative fits to the extinction curve have been presented by Draine and Lee (1984). Désert et al. (1990) have developed a simple yet effective model with three schematic dust components (PAHs, very small grains [VSG] and big grains [BG]), which directly links observed extinction “features” to infrared emission characteristics. Draine and Li (2007) have extended the earlier extinction model of Draine and Lee (1984) to account for the observed IR emission characteristics of the ISM and this has led to some modification of the derived grain size distribution. An alternative model has been presented by Zubko et al. (2004).

In examining these results, it should be realized that there are few constraints from extinction on the grain size distribution at the small end, because for small grains ($<200\text{ \AA}$) extinction is in the Rayleigh limit and independent of grain size (cf. 10.7) and, moreover, the far-UV extinction is dominated by the absorption properties of the material (e.g., ϵ_2 rises rapidly due to electronic transitions) rather than size (cf. 10.7). For very large grains ($\approx 1\text{ }\mu\text{m}$), extinction in the visible is gray (e.g., extinction per unit volume scales with a^{-1}) and dust abundances are mainly derived from abundance constraints on the elements making up these grains. That sets

the near-exponential cutoff of the grain size distribution in [Fig. 10-8](#). Moreover, as emphasized by Zubko et al. (2004), these models are nonunique and the details are dependent on the assumptions.

As alluded to before, mid-IR studies provide an independent handle on the interstellar grain size distribution at the small end. Because of their limited heat capacity, very small species fluctuate in temperature upon absorption of a single UV photon (cf. [Sect. 4.1.2](#)). These temperature fluctuations result in bright mid-IR emission from the ISM ([Fig. 10-4](#)). Sizes of the emitting species (PAHs, PAH clusters, very small grains, and small grains) and typical emission temperatures (cf. [Sect. 2.2](#)) are then approximately linked through the heat capacity in a simple expression, $T_{em} \simeq 3/4 T_m$ where T_m is given by ([10.29](#)). Relative abundances of these components can then be derived once it is recognized that these components compete with “classical” dust grains (responsible for the far-IR continuum emission) for the incident UV photons. The fraction of carbon, f_C , in these compounds is then linked to the ratio of the flux due to this component to the far-IR dust continuum, $f_{IR,i}$, and the dust properties in the UV; namely,

$$f_C = \frac{A_v}{N_H} \frac{\kappa_{uv}}{\kappa_v} \frac{(1 - \omega_{uv})}{\sigma_{uv} A_C} \frac{f_{IR,i}}{(1 - f_{IR,i})} \simeq 0.23 \frac{f_{IR,i}}{(1 - f_{IR,i})}, \quad (10.10)$$

where A_v/N_H is given by ([10.3](#)), the ratio of the UV to visual dust opacity is $\kappa_{uv}/\kappa_v \simeq 3$, the dust albedo $\omega_{uv} \simeq 0.6$, the carbon elemental abundance $A_C = 3.9 \times 10^{-4}$, and the UV opacity $\sigma_{uv} \simeq 7 \times 10^{-18} \text{ cm}^2$ per C-atom for these species.

[Figure 10-9](#) shows the size distribution of these species derived from such a simple analysis. The studies mentioned previously (Désert et al. 1990; Draine and Li 2007; Zubko et al. 2004) are more sophisticated, more self-consistent in absorption and emission, and more precise, but not necessarily more accurate. In any case, the interstellar grain size distribution extends well into the molecular domain. While such a combined size distribution might be fortuitous, it may also reflect an intrinsic relationship between these molecules and dust grains. PAH molecules may be the building blocks of larger dust grains either through chemical growth or through

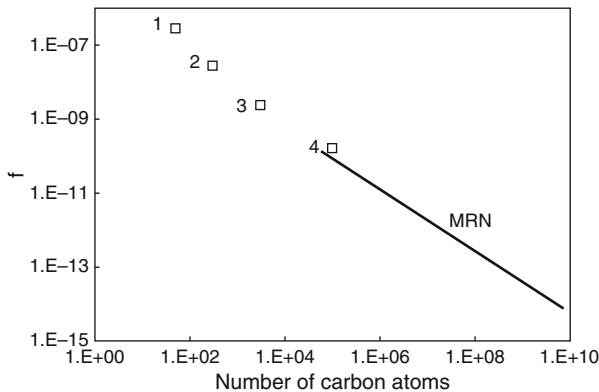


Fig. 10-9

The size distribution of PAHs (1), PAH-clusters (2), very small grains (3; $25 \mu\text{m}$ cirrus), and small grains (4; $60 \mu\text{m}$ grains) derived from IR observations: f is the total number of carbon atoms (per H-atom) locked up in species/grains containing a given number of carbon atoms. For comparison, the MRN grain size distribution derived from extinction studies is also shown

coagulation. Alternatively, PAH molecules may be the shattered fragments produced by the collisional cratering process of large dust grains in interstellar shocks (Cherchneff et al. 1992; Frenklach and Feigelson 1989; Jones et al. 1996).

3.3 Composition

Direct information on the composition of interstellar dust is mainly derived from infrared spectroscopy. The broad 9.7 and 18 μm bands dominate the infrared extinction (► Fig. 10-3). These bands are due to the stretching and bending vibrations of amorphous silicates. The peak strength of the 9.7 μm band is $\tau(9.7)/N_H = (3.5 \times 10^{22})^{-1} \text{ cm}^2$ in the Solar neighborhood (derived from $A_V/\tau(9.7) = 18.5$). With the observed width of 220 cm^{-1} , and a typical integrated absorption strength for silicates of $1.2 \times 10^{-16} \text{ cm (Si-atom)}^{-1}$, this corresponds to an abundance of silicon in solid form of $5.2 \times 10^{-5} \text{ H-atom}^{-1}$. This is almost 50% more than the Solar elemental abundance of silicon and the origin of this discrepancy is not well understood. With an adopted silicate composition of $(\text{Mg,Fe})_2\text{SiO}_4$, the total silicate dust mass is $9 \times 10^{-3} \text{ H-atom}^{-1}$. The silicate abundance is known to vary in the galaxy. Toward the galactic center, the measured $A_V/\tau(9.7)$ is about a factor of 2 less than in the solar neighborhood (A_V is actually scaled from the observed near-IR color excess).

Hydrogenated amorphous carbon (HAC) is the second dust component evident from spectral structure in the infrared extinction curve (► Fig. 10-3) through its aliphatic (e.g., CH_2 and CH_3 groups in sp^3 bonded carbon) CH stretching and bending modes at 3.4, 6.8 and 7.2 μm , respectively. Laboratory spectra of HAC grains provide a good fit to the observed profiles and relative strength of these features. The observed strength of the 3.4 μm feature requires a carbon abundance of about 1.2×10^{-4} relative to H in this dust component or about 1.4×10^{-3} by mass; but keep in mind that only the aliphatic dust component is measured this way and much carbon could be in aromatic compounds. There is no evidence for O atoms in their molecular structure ($\text{O}/\text{H} < 3 \times 10^{-6}$ by number).

Interstellar ice is the third dust component with conspicuous features in the infrared extinction curve (► Fig. 10-3) but these features are only present for sight lines that pass through the dense and shielded environment of molecular clouds. H_2O , NH_3 , CH_4 , CH_3OH , H_2CO , CO , CO_2 , and OCS molecules in an icy matrix have been identified in these environments. The dominant species are H_2O , CO , and CO_2 with abundances of 10^{-4} , 2.5×10^{-5} , and 2.1×10^{-5} relative to H nuclei, relatively. Note, however, that these molecules are not well mixed in the ice. In fact, there are multiple ice components corresponding to either separate layers, or different regions along the line of sight, or mantles on different grain types (Boogert and Ehrenfreund 2004).

Graphite is the final interstellar dust component to be discussed. The presence of graphite in the ISM has been deduced from the strong 2,175 Å feature that dominates the observed interstellar extinction curve in the far-UV (► Fig. 10-2). Graphite has a strong resonance at about 2,200 Å due to $\pi \rightarrow \pi^*$ transitions associated with its aromatic bonds (e.g., sp^2 bonded carbon). Because this is an inherently strong resonance, the exact peak position and profile is very sensitive to grain size, shape, the presence of coatings, and exact optical constants (cf. ► Sect. 3.1 and ► Fig. 10-7). The observed profile of the interstellar 2,175 Å feature is well fitted by theoretically calculated extinction cross sections for either 200 Å graphite spheres or for graphite prolate spheroids with a size of 30 Å and an axial ratio of 1.6 (Draine and Lee 1984). Good fits to the observed spectra can also be obtained with laboratory measured extinction spectra of small (≈ 100 Å) hydrogenated amorphous carbon grains – containing 33% H atoms by

number – which share the aromatic character of their bonding with graphite (Schnaiter et al. 1998). Finally, the 2,175 Å feature is also sometimes ascribed to the $\pi \rightarrow \pi^*$ transition in PAHs and PAH clusters (Steglich et al. 2010). The nonuniqueness implicated by these different suggestions illustrates the sensitivity of the absorption profile to the detailed characteristics of the grain (cf. ▶ Sect. 3.1) or conversely that by adjusting the detailed parameters of the grain slightly, different (but related) materials can be made to fit the observations. Of course, this sensitivity poses a problem for all of these fits in view of the observed constancy of the peak position but variable width of the interstellar 2,175 Å feature; an issue which has not been satisfactorily resolved.

The observed strength of the 2,175 Å feature corresponds to a required abundance of carbon in the form of graphite of 6×10^{-5} relative to H. This corresponds to a mass fraction of graphite of 7.2×10^{-4} relative to H. For HAC grains, the inferred abundance and mass fractions are a factor 2 larger.

Many other dust components have been identified in infrared spectra of dust formed in stellar ejecta or from the analysis of stardust isolated from meteorites. However, their infrared signatures are not present in the diffuse interstellar medium and, typically, that implies a dust abundance in the diffuse ISM less than a few percent. ▶ Table 10-2 summarizes the various compounds identified. Each of the entries in this table is a story in itself, but space does not allow a detailed discussion. Instead the interested reader is referred to various reviews (Tielens 2001; Zinner 2003), which summarize these discussions and provide references to the original literature. Astronomical identifications which are particularly ambiguous are labeled with (?) in this table.

The most striking aspect of ▶ Table 10-2 is the diversity of dust compounds observed. Some 20 dust materials are present in the interstellar/circumstellar dust or stardust record. This diversity reflects to a large extent the heterogeneity of stellar sources contributing to the dust in the interstellar medium. Together these stellar sources represent a wide range in physical conditions (temperatures, pressures, elemental abundances) and, concomitant, dust compounds that can form (cf. ▶ Sect. 5.2). Asymptotic Giant Branch (AGB) stars and type II supernovae are particularly rich in detected dust compounds.

For AGB stars, the astronomical and meteoritic stardust record are in good agreement in the types of dust present. It should be emphasized that most individual AGB objects only show a very limited number of dust compounds and that the diversity of dust reported in ▶ Table 10-2 refers to AGB stars as a class. In particular, the C-rich dust compounds are generally limited to AGB objects whose photosphere is carbon-rich (e.g., elemental C/O > 1), while the oxides and silicates occur generally in oxygen-rich objects (e.g., elemental C/O < 1). Some sources with mixed chemistry exist but this is thought to reflect a recent change over from O-rich to C-rich photosphere or storage of previously ejected material in a long-lived circumstellar disk before the star changed its elemental composition through dredge-up processes. Finally, there is indirect astronomical evidence for a small amount of mass loss during the red giant branch phase, preceding the Asymptotic Giant Branch phase. The composition and amount of dust formed in these outflows is unknown. Meteoritic studies suggest that many of the presolar Al₂O₃ grains are formed in red giant ejecta (Nittler et al. 1997).

In contrast to the observed richness of the AGB objects, the astronomical record of type II SNe is very limited. The best studied, recent supernova, SN 1987A in the Large Magellanic Cloud, showed a featureless mid-IR continuum spectrum, consistent with emission from optically thick clumps (Wooden et al. 2008) in the ejecta. The nearby bright SN, Cas A, showed mid-IR emission which seemed to originate from the fast moving, highly enriched knots and

Table 10-2
An inventory of circumstellar dust

Material	AGB	Post-AGB	PN	Nova	T Tauri	Herbig AeBe	RSG	Wolf Rayet	LBV	SN type II	Massive YSO
Amorphous silicates	1, 2	1	1	1	1	1	1		1	1	1
Crystalline forsterite	1, 2	1	1		1	1	1		1		
Crystalline enstatite	1, 2	1	1		1	1	1		1		
Aluminum oxide	1 (?), 2									2	
Spinel	1 (?), 2									2	
TiO ₂	2										
Hilbonite	2										
MgO	1										
Fe	1 (?)									1	
PAHs	1, 2	1	1	1		1	1	(1)	1		
Amorphous carbon	1	1	1	1				1			
Graphite	2			2						2	
Diamond		1				1				2	
SiC	1, 2		1	2						2	
Other carbides	2	1 (?)								2	
Si ₃ N ₄										2	
MgS	1	1	1						1		
Carbonate			1 (?)		1 (?)	1 (?)					
Ice	1	1	1	1	1	1	1				1

Legend: 1: Astronomical data 2: Meteoritic data

AGB: Low mass (<8 M_⊙) stars on the Asymptotic Giant Branch. Post-AGB object: Low mass (<8 M_⊙) stars in transition from the AGB phase to the Planetary nebula phase. PN: The white dwarf remaining after the phase of prodigious mass loss on the AGB ionizes the AGB-ejecta. The resulting glowing nebula is called a planetary nebula. Nova: The cataclysmic nuclear explosion caused by the accretion of hydrogen onto the surface of a white dwarf star can lead to ejection of material. T Tauri star: A low mass (~1 M_⊙ protostar). Herbig AeBe star: Intermediate-mass (1.5 < M < 10 M_⊙) pre main-sequence stars with spectral types A or B first recognized by Herbig. RSG (Red supergiant): Late and cool (T ~ 3,000 K) stage in the evolution of massive stars (M > 8 M_⊙). Wolf Rayet star: Hot stars characterized by massive stellar winds. Some of these objects have C-rich composition and carbon dust condenses out in their ejecta. LBV (Luminous blue variable): The most massive, brightest, and bluest stars are variable and may experience periods of eruptive mass loss (e.g., η Car). SN type II: The explosion of a massive (M > 8 M_⊙) star at the end of its lifetime. Massive YSO: Luminous and massive protostar characterized by vast amounts of cold dust and gas

a spectrum consistent with amorphous silicates as well as some other dust compounds (Rho et al. 2008). Other, more distant SNe, also show spectral evidence for these materials in their remnants (Rho et al. 2009). The meteoritic record suggests that dust production by supernovae is very diverse but whether this reflects a sampling of a large variety of SNe with very different conditions or the large diversity of chemical zones in each SN remains to be seen. There is no direct astronomical or meteoritic stardust evidence for dust formation in the ejecta of SN type Ia (originating from low mass progenitors). However, such SNe may be a dominant source of iron in the interstellar medium. Because Fe is observed to be highly depleted in the ISM, it is likely that this iron is injected in the form of (nickel)iron grains. Finally, any grains formed in supernovae will have to survive the reverse shock that propagates through the ejecta before it merges with the ISM either because they are protected in dense clumps which are shocked to much lower velocities or because they are very big (grains typically loose $\sim 1,000 \text{ \AA}$ during the adiabatic expansion phase of the supernova remnant).

4 Physics of Interstellar PAHs and Dust

4.1 Temperature and Infrared Emission

4.1.1 Radiative Temperature

The radiative energy balance of a dust grain with geometric cross section, σ_d , is given by

$$4\pi\sigma_d \int_0^\infty Q_{\text{abs}}(\lambda) J(\lambda) d\lambda = 4\pi\sigma_d \int_0^\infty Q_{\text{abs}}(\lambda) B(T_d, \lambda) d\lambda, \quad (10.11)$$

with $J(\lambda)$ the mean intensity of the radiation field at wavelength λ , T_d the radiative equilibrium dust temperature, and $B(T_d, \lambda)$ the Planck function. For the diffuse interstellar medium, the integrated mean intensity is equivalent to ($4\pi J = cU$ with U the energy density) about $7 \times 10^{-13} \text{ erg cm}^{-3}$ for stellar light and $4 \times 10^{-13} \text{ erg cm}^{-3}$ for the 3 K background radiation. For a black body, the radiative temperature is then $\simeq 3.5 \text{ K}$. However, small dust grains are not black bodies. Often, for simplicity, the absorption efficiency is represented by

$$Q(\lambda) = Q_o \left(\frac{\lambda_o}{\lambda} \right)^\beta \quad \text{for } \lambda > \lambda_o, \quad (10.12)$$

with Q_o and β constants and λ_o a reference wavelength. Typically, β 's between 1 and 2 are considered. For $\beta = 1$, this expression is sometimes further simplified by setting $Q_o = 1$ and $\lambda_o = 2\pi a$. This leads to

$$T_d = \left(\frac{hc}{k} \right) \left(\frac{4\pi J}{384\pi^2 a hc^2 \zeta(5)} \right)^{1/5}, \quad (10.13)$$

with ζ the Riemann Zeta function ($\zeta(5) = 1.037$). Such a grain heated by stellar radiation in the diffuse ISM reaches a temperature, $T_d \simeq 14 (1,000 \text{ \AA} / a)^{0.2} \text{ K}$. Using realistic dust efficiencies, calculated dust temperatures are well represented by

$$T_{\text{sil}} = 15.6 \left(\frac{1,000 \text{ \AA}}{a} \right)^{0.06} \text{ K}, \quad (10.14)$$

for silicates and,

$$T_{\text{gra}} = 18.1 \left(\frac{1,000 \text{ \AA}}{a} \right)^{0.06} \text{ K}, \quad (10.15)$$

for graphite grains.

Dust grains become much hotter close to luminous stars. The radiative flux at a distance, d , from a star with luminosity, L_* , can be expressed in terms of the average radiation field of the diffuse ISM (the Habing field, $1.6 \times 10^{-3} \text{ erg cm}^{-2} \text{ s}^{-1}$) as

$$G_o = 2.1 \times 10^4 \left(\frac{L_*}{10^4 L_\odot} \right) \left(\frac{0.1 \text{ pc}}{d} \right)^2. \quad (10.16)$$

Taking into account that, in this one-dimensional radiation field, absorption scales with the geometric cross section while emission occurs with the actual surface area (πa^2 vs. $4\pi a^2$ for spherical grains), the dust temperature of a fiducial grain with $\beta = 1$ is given by

$$T_d \simeq 53 \left(\frac{1,000 \text{ \AA}}{a} \right)^{0.2} \left(\frac{G_o}{10^4} \right)^{0.2} \text{ K}, \quad (10.17)$$

for silicates,

$$T_{\text{sil}} \simeq 57 \left(\frac{1,000 \text{ \AA}}{a} \right)^{0.06} \left(\frac{G_o}{10^4} \right)^{1/6} \text{ K for } T_{\text{sil}} < 250 \text{ K}, \quad (10.18)$$

and for graphite,

$$T_{\text{gra}} \simeq 70 \left(\frac{1,000 \text{ \AA}}{a} \right)^{0.06} \left(\frac{G_o}{10^4} \right)^{1/5.8} \text{ K for } T_{\text{gra}} < 70 \text{ K}. \quad (10.19)$$

The IR intensity is given by

$$I(\lambda) = \int n_d(\ell) \pi a^2 Q(\lambda) B(T_d, \lambda) d\ell, \quad (10.20)$$

where n_d is the dust density and the integral is evaluated along the line of sight, ℓ . For a homogenous cloud of size L , this becomes

$$I(\lambda) = B(T_d, \lambda) \tau_d(\lambda), \quad (10.21)$$

with the optical depth given by (10.6). For a grain size distribution, $n_d(a)$, the intensity equation is

$$I(\lambda) = \int \int_{a_-}^{a_+} n_H(\ell) n(a) \pi a^2 Q(\lambda, a) B(T_d, \lambda) da d\ell, \quad (10.22)$$

where the size distribution, $n(a)$, is given by (10.9) and n_H is the H-nuclei density.

4.1.2 Temperature Fluctuations

The expressions derived for the dust temperature in Sect. 4.1.1 assume a balance between emission and absorption. However, for a small species with a limited heat capacity, these two processes can operate on very different timescales and the temperature immediately after UV photon absorption – and during the emission process – is much higher than the average temperature. The temperature follows then from the time-dependent evolution of the energy equation.

Here, this will be evaluated for a single photon event that heats the species to a high temperature from where it cools down through IR vibrational emission.

In the analysis, a PAH molecule with, N_c , carbon atoms is represented by a disk with a size of

$$a \simeq 0.9 \times 10^{-8} N_c^{1/2} \text{ cm}, \quad (10.23)$$

and a total surface area,

$$\sigma_{\text{PAH}} \simeq 2\pi a^2 \simeq 5 \times 10^{-16} N_c \text{ cm}^2. \quad (10.24)$$

The timescale for UV absorption is

$$\tau_{\text{uv}} = k_{\text{uv}}^{-1} = (4\pi \sigma_{\text{uv}}(\text{PAH}) \mathcal{N}_{\text{uv}})^{-1} \simeq \frac{1.4 \times 10^9}{N_c G_o} \text{ s}, \quad (10.25)$$

with $\sigma_{\text{uv}}(\text{PAH})$ the UV absorption cross section of the PAH (approximately equal to $7 \times 10^{-18} \text{ cm}^2$ per carbon atom (N_c)) and \mathcal{N}_{uv} the mean photon intensity of the radiation field, here expressed in terms of the Habing field, G_o ($\simeq 10^8 \text{ photons cm}^{-2} \text{ s}^{-1}$). So, a typical interstellar PAH with $N_c = 50$ absorbs a UV photon once a year in the diffuse ISM and every 10 min in a PDR such as the Orion Bar. The electronic excitation energy is rapidly interconverted into vibrational excitation. The vibrational excitation energy is then radiated away on a timescale, $\sim 1 \text{ s}$. So, the PAH will be very hot $\sim 1,000 \text{ K}$ immediately after photon absorption and then cool down to “background” temperatures until the next absorption occurs.

Thus, for the “emission” spectrum of a small species, the internal energy, E , has to be linked to the excitation temperature of the vibrational modes through the heat capacity or equivalently the density of states. Consider a single mode, i , in a PAH connected to a thermal bath represented by all other modes. Under certain limitations, the system can then be considered a canonical ensemble and the excitation of this mode can be described by a temperature, T , equal to the mean energy in the mode in units of the Boltzmann constant, k . Of course, a truly canonical ensemble requires that all energies are accessible and that is not the case here since the mode can never have more energy than the total energy, E , in the system. In practice, however, the canonical approach leads to reasonable results as long as the average energy in mode, i , is small compared to the total energy in the system. Or, more generally, for a system to be a canonical heat reservoir, its internal temperature should not be affected by the (small) amount of energy exchanged with the mode under consideration. Because the energy will fluctuate around the mean energy and, because the vibrational excitation is so heavily weighted toward higher energies, the average excitation will be overestimated in the canonical approximation. For a species with 100 C-atoms, this results in a factor 2 error in the fractional excitation when the energy in the system is $\simeq 4$ times the energy of the mode. The error decreases to a factor 1.1 when the energy in the system increases to ~ 10 times the energy of the mode. Thus, the error is largest for the highest frequency modes.

Introducing the entropy, S , of a species,

$$S \equiv k \ln [\rho(E)], \quad (10.26)$$

with $\rho(E)$ the density of states for a species with internal energy, E , the microcanonical temperature, T_m , is given by

$$\frac{1}{kT_m} = \frac{1}{k} \frac{dS}{dE} = \frac{d \ln [\rho(E)]}{dE}. \quad (10.27)$$

Several methods exist to evaluate the density of states of small PAH molecules. The calculated density of states of PAH molecules can be well approximated by

$$\ln[\rho(E)] = 2.84 \times 10^{-2} s \left(\frac{E}{s}\right)^{0.60} - 6.15 \quad (10.28)$$

with s the number of degrees of freedom ($s = 3N_c - 6$), E in units of cm^{-1} , and $\rho(E)$ in units of states per cm^{-1} . This approximation is valid over the energy range 2.5×10^{-2} to $3 \times 10^2 \text{ cm}^{-1}$ per mode (10^{-5} – 10^{-1} eV per C-atom). The H-atoms at the periphery of the PAH do not contribute much to the internal energy and these are ignored here. The microcanonical temperature is then given by

$$T_m \simeq 2,000 \left(\frac{E(\text{eV})}{N_c}\right)^{0.4} \text{ K}, \quad (10.29)$$

with E in electron volts. This equation is approximately valid over the range 35–1,000 K (within 10%) but rapidly deteriorates outside of these limits. Consider now a 50 C-atom PAH, its “radiative” equilibrium temperature would be ~ 25 K according to (10.19). In contrast, immediately after absorption of a 10 eV photon, the PAH excitation temperature is 1,050 K, but cooling within a few seconds to a very low temperature. For very low internal energies (< 0.1 eV), the concept of temperature is not well defined anymore, since the vibrational modes will have decoupled and excitation energy is “trapped” in specific mode(s) that may even not be radiatively active and, in any case, will decay very slowly.

4.1.3 Temperature Distribution Function and PAH Spectrum

Consider now a small species which absorbs a UV photon, with energy $h\nu$ and reaches an initial temperature, T_i , given by (10.29). Assuming that IR photon emission is a Poisson process, the temperature distribution function, $G(T)$, is given by

$$G(T) dT = \frac{k_{\text{uv}}}{dT/dt} \exp[-k_{\text{uv}} \tau_{\text{min}}(T)] dT, \quad (10.30)$$

where the UV absorption rate is given by (10.25). The time, $\tau_{\text{min}}(T)$, to cool down from the initial temperature, T_i , to T is given by the integral expression,

$$\tau_{\text{min}}(T) = \int_T^{T_i} \frac{1}{dT/dt} dT. \quad (10.31)$$

The cooling function is given by

$$\frac{dE}{dt} = -4\pi \sum_i \kappa_i B(\nu_i, T), \quad (10.32)$$

where κ_i is the absorption coefficient of vibrational mode i and the sum is over all modes. The temperature decay law, dT/dt can then be obtained using the specific heat. In principle, dT/dt will depend on the detailed characteristics of the specific PAH under consideration. A good approximation for small (≈ 50 C-atoms) neutral PAHs is given by

$$\frac{dT}{dt} \simeq -1.1 \times 10^{-5} T^{2.53} \text{ K s}^{-1} \quad T > 250 \text{ K}, \quad (10.33)$$

At a temperature of 1,000 K, this corresponds to a cooling timescale of ≈ 2 s. Note that, to first order, the energy cooling rate is independent of the PAH size because at a given temperature the

emission rate as well as the energy content scale similarly with the number of modes. Of course, for a given absorbed photon energy, a larger PAH will be less excited and hence its cooling rate will be slower. Ionized PAHs are somewhat (factor 3–4) more effective coolers due to the larger intrinsic strength of the C–C modes in such species.

For small PAHs, single photon events are all that counts but for larger species multiphoton events have to be taken into account. Typically, when $\tau_{\text{uv}} < 0.1(dT/dt)^{-1}$, multiphoton processes are important. This translates into, $G_o > 10^6 (50/N_c)^{1.6}$ for an absorbed UV photon energy of 10 eV. In that case, T_o and T_1 have to be determined in an iterative fashion. One way to do this is by iterating on $G(T)$, namely

$$G_{n+1}(T_o, T) = \frac{1}{\bar{\tau}} \int_{T_o}^T G_n(T_o, T') G(T', T) dT' \quad (10.34)$$

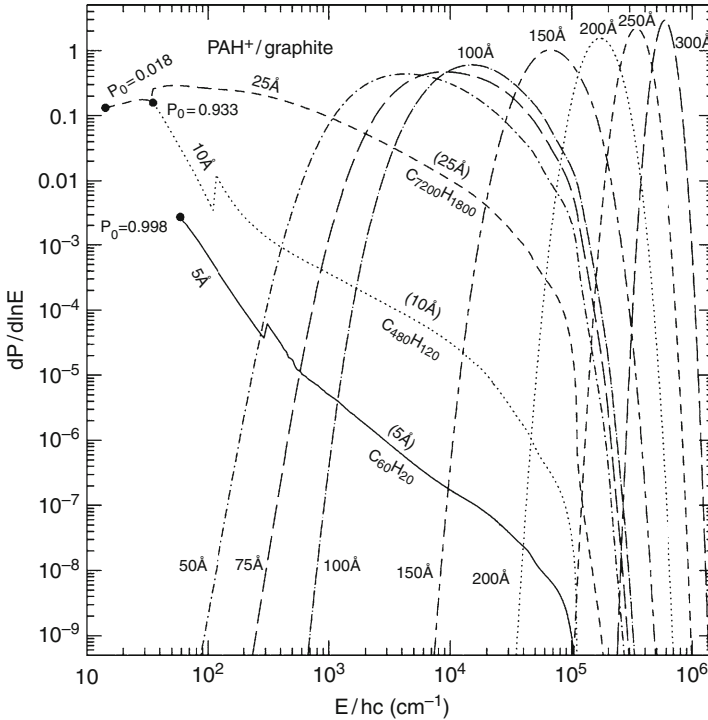
where the term in the integral is the probability that, starting at temperature, T_o , the species has a temperature T' after n photon events multiplied by the probability that the absorption of an additional photon takes it to temperature T . Starting with a delta function at the microwave background temperature, this system can then be iterated until convergence is attained. Details on this and various other methods for the temperature distribution function of small species can be found in Aannestad (1989), Bakes et al. (2001), and Draine and Li (2001).

In a general sense, the photon absorption and emission process can be treated as a stochastic process, even for large grains. Focus first on the absorption process. At the radiative steady state, the amount of energy absorbed equals the amount of energy emitted. Define $\langle h\nu_{\text{uv}} \rangle$ as the average photon energy absorbed and \bar{E} as the average energy of the species (which is related to the temperature through the heat capacity). Then, N_{abs} photons have to be absorbed in order to reach this steady state value with $N_{\text{abs}} = \bar{E} / \langle h\nu_{\text{uv}} \rangle$. For a stochastic process, the dispersion in the number of photons absorbed is then $N_{\text{abs}}^{1/2}$ or in terms of the dispersion in energy, $\sigma_{\text{abs}} = (\bar{E} \langle h\nu_{\text{uv}} \rangle)^{1/2}$. When N_{abs} is large, the distribution function is then given by the gaussian

$$P(E) = \frac{\exp\left(-\frac{(E - \bar{E})^2}{2\bar{E} \langle h\nu_{\text{uv}} \rangle}\right)}{\sqrt{2\pi\bar{E} \langle h\nu_{\text{uv}} \rangle}} \quad (10.35)$$

So, as the average internal energy becomes larger and larger compared to the typical absorbed photon energy, the energy distribution function will become more and more sharply peaked around the mean value. The IR emission process will in the same way give rise to a Gaussian distribution with a dispersion set by the number of photons emitted; e.g., $\sigma_{\text{em}} = (\bar{E} \langle h\nu_{\text{ir}} \rangle)^{1/2}$ with $\langle h\nu_{\text{ir}} \rangle$ the average energy of the IR photons that are emitted. The combined absorption and emission distribution function is then also a Gaussian with a dispersion given by $\sigma^2 = \sigma_{\text{uv}}^2 + \sigma_{\text{ir}}^2$ and because the average IR emission energy is so much less than the UV absorption energy, $\sigma \simeq \sigma_{\text{uv}}$. Calculated energy distribution functions for different sized species are shown in [Fig. 10-10](#). For PAHs, single photon events dominate and the distribution function is given by [Fig. 10.30](#). As the species increases in size, the distribution function transits to a Gaussian distribution and when $\bar{E} \gg h\nu_{\text{uv}}$ becomes very sharply peaked at \bar{E} .

The temperature distribution of interstellar PAHs will thus depend on the size of the species. This is illustrated in [Fig. 10-11](#). Classical-sized grains are largely in radiative equilibrium with their environment in the diffuse ISM. As the size of the species decreases, temperature fluctuations become more prevalent. Very small grains are very cold except directly after UV photon absorption. For PAH molecules the maximum temperature after FUV photon absorption can reach 2,000 K and decays very rapidly (~ 2 s).



■ Fig. 10-10

The internal energy distribution function for different sized species. The kinks in the distribution functions for the smallest sizes are artifacts due to approximations in the adopted properties (Figure taken from Li and Draine 2001)

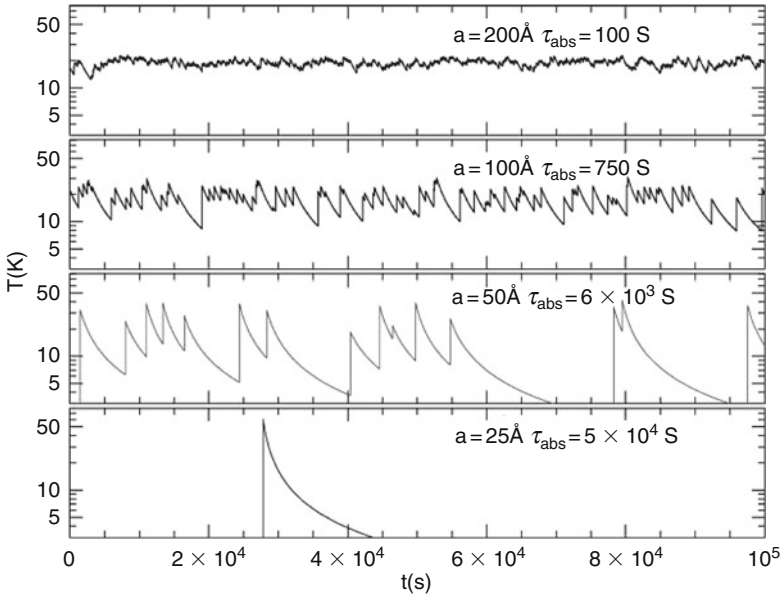
The emission intensity due to the vibrational transition $\nu \rightarrow \nu - 1$ in mode i is then given by

$$I(E, i, \nu) = \frac{N_{\nu, i}(E) A_{\nu, i} \hbar \nu_i}{4\pi}, \quad (10.36)$$

with $A_{\nu, i}$ the Einstein A emission coefficient associated with this transition and $N_{\nu, i}(E)$ all species with internal energy E with ν quanta in mode i along the line of sight. Thus, using (10.29) to describe the excitation of a PAH with internal energy, E , the intensity can be calculated. More sophisticated methods take the full energy cascade into account as the species cools down due to IR photon emission, using for example a Monte Carlo technique. For harmonic oscillators, this equation transforms to the usual intensity relation for optically thin emission,

$$I(E, i) = \tau_i B(\nu_i, T), \quad (10.37)$$

with $B(\nu_i, T)$ the Planck function and τ_i the optical depth in mode i . This is a good approximation for global spectral calculations but, when evaluating detailed spectral characteristics – in particular, profiles – it should be kept in mind that the PAH vibrational modes are highly anharmonic. In this, the discussion on the difference between canonical and microcanonical ensembles discussed above should also be kept in mind.



■ Fig. 10-11

The time-dependent behavior of the temperature for various sizes of the species. The time axis corresponds to approximately a day. These calculations pertain to the diffuse ISM (Figure taken from Draine 2003)


In reality, an interstellar PAH will be exposed to a spectrum of exciting UV photons. This set of equations should then be averaged over the UV spectrum; e.g., each UV absorption frequency yields a distinct T_1 and the temperature distribution function, $G(T)$, has to be properly averaged over this range. Of course, the rate of UV photon absorption (◆ 10.25) has to be integrated over the excitation spectrum as well. Finally, the interstellar PAH family may be very diverse and occur in different ionization stages, each with its own intrinsic properties, including distinct vibrational modes. The IR intensity of such a family can then be found by a proper summation, namely

$$I(\nu) = \sum_k \sum_j \sum_i n(k) f(j) I_{k,j}(i) \phi_{k,j}(i, \nu), \quad (10.38)$$

where the summation is over the modes, i , of ionization stage, j , of molecule, k . Here, $n(k)$ is the density of the specific PAH molecule, k , and $f(j)$ the fractional abundance of ionization stage, j . The integrated intensity, $I_{k,j}(i)$, of mode i of molecule k in ionization stage j is given by (◆ 10.37), and $\phi_{k,j}(i, \nu)$ is the intrinsic line profile of this transition.

4.1.4 IR Emission Models

Detailed dust models have been developed to model the infrared emission spectrum of the interstellar medium. Often, these models are based upon general dust models; e.g., they assume a few specific materials to be present (e.g., graphite, silicates), adopt appropriate (or somewhat

modified) optical constants, “invert” the observed extinction curve to derive a grain size distribution that is consistent with the observations. This model can then be used to calculate the IR emission spectrum. Variations in observed emission spectra can then be interpreted in terms of variations in the size distribution. The most-insightful analysis of this has been presented by Désert et al. (1990). This model realized that extinction studies can be summarized by identifying three independent dust components responsible for the visible and near-ultraviolet linear, the 2,175 Å bump, and the nonlinear far-ultraviolet rise, respectively. Similarly, the infrared emission spectrum contains three independent components: the mid-infrared emission features, the 25 and 60 μm cirrus emission due to fluctuating grains, and the far-infrared emission due to large grains in radiative equilibrium with the radiation field. This model then links the extinction and emission components – using laboratory data and common sense as a guide – and derived a very heuristic dust model. These three components have been “baptized” PAH molecules, Very Small Grains, and Big Grains. A fit to the observed galactic cirrus is shown in  Fig. 10-4. Basically, the peak position of the far-infrared emission sets the temperature of the Big Grains and hence the strength of the radiation field (cf. [10.17–10.19]). With the temperature, the column density of Big Grains is then also known. The PAHs spectrum is independent of G_o and the PAH abundance relative to that of the Big Grains is set by the level of mid-IR features relative to the far-IR. The spectrum of the Very Small Grains is somewhat sensitive to G_o since multiphoton events are important. However, this can be accounted for and the strength of the mid-IR continuum provides then the abundance of Very Small Grains. Hence, observed variations in the spectral energy distribution can be linked to the key parameters describing the dust: the relative abundance of these three components. While more complex models have been derived (Draine and Li 2007; Galliano et al. 2005), in essence they all boil down to this simple prescription. When applying this to realistic sources (e.g., galaxies), variations in G_o may have to be taken into account, but that is a minor modification. If there is an estimate of the gas mass from HI studies, the dust-to-gas ratio can also be derived and compared, for example, to the metallicity of the galaxy. Both ISO and Spitzer data on galaxies has been analyzed wholesale this way and, while details vary between the different models – a sign that systematics play a role – trends in the spectra can be analyzed in a coherent and meaningful way (Draine et al. 2007; Galliano et al. 2005).

4.2 Charge

There are three main processes that set the charge of a grain: the photoelectric effect, and electron and positive ion recombination. Generally, it is sufficient to take singly charged ion collisions into account and the charge balance equation reads (Bakes and Tielens 1994; Draine and Sutin 1987),

$$f(Z_d) [J_{pe}(Z_d) + J_{ion}(Z_d)] = f(Z_d + 1) J_e(Z_d + 1), \quad (10.39)$$

with $f(Z_d)$ the probability of finding a grain at charge $Z_d e$, J_{pe} the rate of photo electron emission, and J_e and J_{ion} the accretion rates of electrons and ions, respectively. The charge distribution functions follows then from successive applications of the following two sets of equations,

$$f(Z_d) = f(0) \prod_{Z'_d=1}^{Z_d} \left[\frac{J_{pe}(Z'_d-1) + J_{ion}(Z'_d-1)}{J_e(Z'_d)} \right] \quad Z_d > 0 \quad (10.40)$$

$$f(Z_d) = f(0) \prod_{Z'_d=Z_d}^{-1} \left[\frac{J_e(Z'_d+1)}{J_{pe}(Z'_d) + J_{ion}(Z'_d)} \right] \quad Z_d < 0, \quad (10.41)$$

closed by

$$\sum_{Z_d=-\infty}^{\infty} f(Z_d) = 1. \quad (10.42)$$

So, in order to solve the charge balance of the dust, the photoelectric emission rate and the electron and ion collision rates have to be specified.

4.2.1 Collisional Rates

The collisional rate, $J_i(Z_d)$, of gas particles, i , with charge q_i , number density, n_i , mass, m_i , with a grain of charge $Z_d e$ is given by

$$J_i(Z_d) = n_i s_i \left(\frac{8kT}{\pi m_i} \right)^{1/2} \pi a^2 \tilde{J}(\tau, \nu), \quad (10.43)$$

where the reduced temperature, $\tau = akT/q_i^2$, and the charge ratio, $\nu = Z_d e/q_i$ have been introduced. Here, s_i is an effective sticking coefficient and $\tilde{J}(\tau, \nu)$ is the reduced rate which contains the Coulomb interaction aspects averaged over the velocity distribution (Draine and Sutin 1987). At low kinetic energies of the impacting species, the sticking coefficient is likely near unity, even for an electron. The sticking coefficient may be reduced at higher energies due to secondary electron emission and because the impacting species may actually traverse a small grain and exit on the other side.

Assuming a Maxwellian velocity distribution and that the charge interacts through the image potential, the reduced rate for neutral grains ($\nu = 0$) is given by

$$\tilde{J}(\tau, \nu = 0) = 1 + \left(\frac{\pi}{2\tau} \right)^{1/2}. \quad (10.44)$$

For attractive interaction ($\nu < 0$), the cross section is enhanced by electrostatic focussing. The reduced rate is then to a good approximation,

$$\tilde{J}(\tau, \nu < 0) \simeq \left(1 + \frac{|\nu|}{\tau} \right) \left(1 + \left(\frac{2}{\tau + 2|\nu|} \right)^{1/2} \right), \quad (10.45)$$

where the first term on the right-hand side is the familiar Coulomb focusing factor ($1 + |Z_d e q_i|/akT$) and the second factor represents the image polarization interaction. For repulsive interaction ($\nu > 0$), a good approximation is provided by

$$\tilde{J}(\tau, \nu > 0) \simeq \left(1 + (4\tau + 3\nu)^{-1/2} \right)^2 \exp \left[\frac{-\theta_\nu}{\tau} \right], \quad (10.46)$$

with

$$\theta_\nu \simeq \frac{\nu}{1 + \nu^{-1/2}} \quad (\nu > 0). \quad (10.47)$$

The exponential term in this expression reflects that only those collisions which have enough initial kinetic energy (in the Maxwellian distribution) to overcome the repulsive interaction potential will contribute.

The electron recombination rate is given by

$$J_i(Z_d) = 2 \times 10^{-3} \left(\frac{T}{100 \text{ K}} \right)^{1/2} n_e \left(\frac{a}{1,000 \text{ \AA}} \right)^2 \tilde{J}(\tau, \nu) \quad \text{s}^{-1}. \quad (10.48)$$

For diffuse clouds where C^+ is the dominant ion, the ion recombination rate is a factor 6.9×10^{-3} smaller and generally not very important compared to the photoelectric charging.

4.2.2 Photoelectric Rates

The photoelectric ejection rate is given by Bakes and Tielens (1994)

$$J_{pe}(Z_d) = 4\pi \int_{\nu_{Z_d}}^{\nu_H} \frac{J(\nu)}{h\nu} \sigma_d(\nu) Y_{\text{ion}}(Z_d, \nu) d\nu. \quad (10.49)$$

The photoionization yield, Y_{ion} , initially rises rapidly with photon energy above the ionization potential and then levels off at a constant yield for higher energies. A semiempirical relation for this yield is given by

$$Y_{\text{ion}}(Z_d, \nu) = Y_{\infty} \left(1 - \frac{IP(Z_d)}{h\nu} \right) f_y(a), \quad (10.50)$$

where Y_{∞} is the photoionization yield for bulk materials and $f_y(a)$ is the yield enhancement factor for small grains. While for bulk materials, the ionization potential, IP , is equal to the workfunction, W (4–6 eV for relevant materials), this is not the case for very small grains. For small grains, the ionization potential increases because of the electrostatic work required to charge up the particle,

$$IP(Z_d) - W = \left(Z_d + \frac{1}{2} \right) \frac{e^2}{C}. \quad (10.51)$$

The capacitance, C , is a for a spherical grain and $C = 2a/\pi$ for a disk (e.g., a PAH molecule). For small ($a \simeq 3|Z_d| \text{ \AA}$) grains and PAHs, the ionization potential can become rapidly very large, severely limiting the total charging up such a grain may experience. In particular, in a diffuse cloud, where electron recombination is counteracted by ionization with photons less than 13.6 eV, a 50 C-atom PAH can only be charged to $Z \simeq 3$ (with $W = 4.4 \text{ eV}$).

The yield enhancement factor, $f_y(a)$, takes into account that for bulk materials the photon attenuation depth, l_a , in the material is larger than the electron mean free path, l_e . Photoelectrons, created deep within bulk material, never reach the surface and the photoelectric effect is suppressed. This factor is approximately given by

$$f_y(a) = \left(\frac{\zeta}{\alpha} \right)^2 \frac{(\alpha^2 - 2\alpha + 2 - 2 \exp[-\alpha])}{(\zeta^2 - 2\zeta + 2 - 2 \exp[-\zeta])}, \quad (10.52)$$

with $\alpha = a(l_e + l_a)/l_e l_a$ and $\zeta = a/l_a$. For small grains, f_y is given by $(l_e + l_a)/l_e$, while for large grains, f_y goes to unity (e.g., the photoelectric yield is normalized to that measured for bulk materials and is enhanced for small grains). Typical values for Y_{∞} , and l_e and l_a are 0.15, 10 and 100 \AA , respectively.

Assuming absorption cross sections for very small graphitic grains and the diffuse interstellar radiation field, the photoelectric ionization rate can now be found from (\blacktriangleright 10.49),

$$J_{pe}(Z_d) \simeq 2.5 \times 10^{-13} (13.6 - IP(Z_d))^2 N_c f_y(a) G_o \quad \text{electrons s}^{-1}, \quad (10.53)$$

where it is assumed that the total FUV absorption cross section scales with the total number of carbon atoms, N_c , in the grain. For grains larger than $a > 100 \text{ \AA}$, the FUV absorption cross section will scale with the surface area,

$$J_{pe}(Z_d) \simeq 1.6 \times 10^{-7} (13.6 - IP(Z_d))^2 \left(\frac{a}{100 \text{ \AA}}\right)^2 f_y(a) G_o \quad \text{electrons s}^{-1}. \quad (10.54)$$

Setting the ionization potential equal to the workfunction ($\simeq 5 \text{ eV}$) and f_y to unity, J_{pe} is approximately equal to $2.2 \times 10^{-5} (a/100 \text{ \AA})^2 G_o \text{ electrons s}^{-1}$.

4.2.3 The Charge Distribution Function

In most of the diffuse interstellar medium, the grain charge is dominated by the photoelectric effect balanced by electron recombination (Bakes and Tielens 1994). The abundance ratio of adjacent ionization stages is then given by

$$\frac{f(Z_d + 1)}{f(Z_d)} = \frac{J_{pe}(Z_d)}{J_e(Z_d + 1)}, \quad (10.55)$$

where Z_d is of course positive and ion-grain recombination has been neglected. For large, multiply charged grains in the diffuse ISM with an ionization potential equal to the work function ($\simeq 5 \text{ eV}$), this evaluates to

$$\frac{f(Z_d + 1)}{f(Z_d)} \simeq 6.1 \frac{G_o}{n_e T^{1/2}} \left[1 + \frac{Z_d + 1}{\tau}\right]^{-1}. \quad (10.56)$$

The Coulomb interaction prefers the slowest electrons and, when Coulomb focussing is important ($(Z_d + 1)/\tau$ is large), the ionization parameter $\gamma = G_o T^{1/2}/n_e$ describes the ionization distribution.

The peak of the distribution, Z_p , occurs when electron ejection and electron recombination balance,

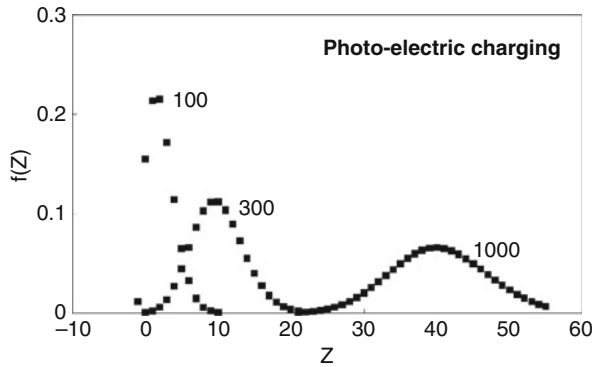
$$Z_p \simeq 36 \left(\frac{\gamma}{1,000 \text{ cm}^3 \text{ K}^{1/2}}\right) \left(\frac{a}{1,000 \text{ \AA}}\right), \quad (10.57)$$

where γ is of the order of $1,000 \text{ cm}^3 \text{ K}^{1/2}$ in the diffuse ISM. Thus, the photoelectric effect can charge up grains considerably in the diffuse ISM with potentials of $Z_p e^2/akT \simeq 6\gamma/T \simeq 60$. Realizing that, statistically speaking, charging is like flipping a coin with a probability of success (increase in positive charge) relative to failure given by the right-hand side of (10.55) – which is only weakly dependent on the actual grain charge when Z_p is large – the distribution function will be a Gaussian with a peak at Z_p and a width given by

$$\sigma_Z = \sqrt{Z_p}. \quad (10.58)$$

Similar considerations apply for HII regions where electron and ion recombinations balance and again a Gaussian is (approximately) obtained. Inside dense clouds, the degree of ionization is very low and, in that case, a binomial distribution is obtained with Z_p between -1 and 0 ; e.g., the grain is part of the time neutral and part of the time negatively charged.

Some representative results for the grain charge distribution function are shown in Fig. 10-12. The charge distribution functions are reasonably well represented by Gaussians. However, as a result of the various approximations made, the peak positions are somewhat displaced from the simple analytical estimates (10.57). As expected, larger grains are more



■ Fig. 10-12

Typical charge distribution function for interstellar grains in diffuse clouds assuming an ionization parameter, γ , of $1,500 \text{ cm}^3 \text{ K}^{1/2}$. The three distributions are labeled by the grain size in angstrom

positively charged and their charge distribution is also broader. The electrostatic grain potential is given by

$$\phi_d = \frac{Z_d e^2}{a}, \quad (10.59)$$

which is typically independent of grain size. In particular, in a photoelectric charging environment, ϕ is approximately $0.5 (\gamma/1,000 \text{ cm}^3 \text{ K}^{1/2}) \text{ eV}$.

5 The Life Cycle of Interstellar PAHs and Dust

The life cycle of interstellar dust starts with the nucleation and growth of high-temperature condensates such as silicates, graphite, and carbides at high densities and temperatures in the ejecta from stars. This ejected material is rapidly mixed with other gas and dust in the interstellar medium (ISM). In the ISM, dust cycles many times between the intercloud and cloud phases on a very fast timescale ($\approx 3 \times 10^7$ year). In the low density, warm neutral and ionized intercloud media, dust is processed by strong shocks driven by supernova explosions. The hot gases in the shock can sputter atoms from the grains. Also, high-velocity collisions among grains can lead to vaporization, melting, phase transformation, and shattering of the projectile and target. In the denser media – diffuse and dense clouds – gas phase species can accrete onto grains forming a mantle. Coagulation may also play a role in increasing the grain size inside diffuse and dense clouds. If the grain survives the onslaught of interstellar shocks, eventually, during one of these cycles, a grain may find itself in a dense cloud core when this core becomes gravitationally unstable against collapse. The grain may then wind up in the star or in the surrounding planet forming disk. The complete cycle from injection by a star until formation of a new star and any associated planets typically takes some 2×10^9 year.

■ **Table 10-3**

Interstellar gas and dust budgets

Source	\dot{M}_H^a ($M_\odot \text{ kpc}^{-2} \text{ M year}^{-1}$)	\dot{M}_c^b ($M_\odot \text{ kpc}^{-2} \text{ M year}^{-1}$)	\dot{M}_{sil}^c ($M_\odot \text{ kpc}^{-2} \text{ M year}^{-1}$)
C-rich giants	750	3	–
O-rich giants	750	–	5
Novae	6	0.3	0.03
SN type Ia	–	0.3 ^d	2 ^d
OB stars	30	–	–
Red supergiants	20	–	0.2
Wolf Rayet	100	0.06 ^e	–
SN type II	100	2 ^d	10 ^d
YSO	(1,500) ^f		8

Taken from Tielens (2001). Except for SN type II,^d these values are uncertain by a factor ~ 2

^aTotal gas mass injection rate

^bCarbon dust injection rate

^cSilicate, oxide, and metal dust injection rate

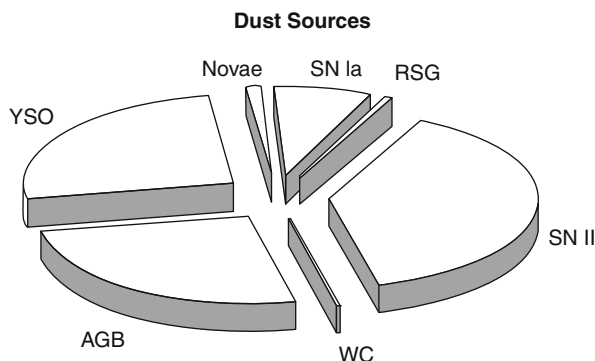
^dFraction and composition of dust formed in SN is presently unknown. These values correspond to upper limits

^eDust injection only by carbon-rich WC 8–10 stars

^fWinds from young stellar objects merely take circumstellar gas and return it to the “general” molecular cloud. This cycling occurs through the disk and dust entrained in these winds may have been modified or even completely vaporized and recondensed in the inner protoplanetary disk

5.1 Sources of Interstellar Dust

A large number of stellar sources contribute to the dust in the interstellar medium (cf. [▶ Tables 10-2](#) and [▶ 10-3](#)). Astronomical studies show that AGB stars are an important source of dust in the interstellar medium ([▶ Fig. 10-13](#)). About half of the AGB-dust-mass is in the form of amorphous carbon dust injected by C-rich AGB stars, while the other half is in the form of silicates injected by O-rich AGB stars. In contrast, other known sources of interstellar dust – Novae, WC Wolf-Rayet stars, red supergiants – contribute only traces of dust. The onset of dust formation in supernova ejecta can be discerned from an increase in the thermal infrared emission accompanied by a decrease in visible SN-light as well as through the development of a blue-red asymmetry in SN emission line profiles. However, the amount of dust formed is very difficult to ascertain (Barlow 2009; Meikle et al. 2007; Sugerman et al. 2006). SN 1987A is known to have formed at least $10^{-4} M_\odot$ of dust and perhaps as much as $0.1 M_\odot$ of dust (Wooden et al. 2008). For about half a dozen other SNe the signature of the onset of dust formation has been detected but with very uncertain dust-mass estimates. A recent submillimeter study suggested that the more evolved SN, Cas A, was very efficient in forming dust, but this was contaminated by foreground interstellar dust. The IR evidence suggests a low dust formation efficiency in Cas A with $0.02\text{--}0.05 M_\odot$ of dust (Barlow 2009; Rho et al. 2008, 2009) for a total ejecta mass of $\approx 4 M_\odot$. Theoretically, dust formation is calculated to be efficient (Kozasa et al. 1984) but those studies – based upon incomplete physical and chemical insight in dust formation as well as a limited understanding of SN explosions – have presently little predictive value. More recent theoretical studies have focused on understanding the kinetics involved in the formation of



■ Fig. 10-13

The contribution of different stellar sources to the interstellar dust budget. Note that all these values are very uncertain (Taken from Tielens 2001)

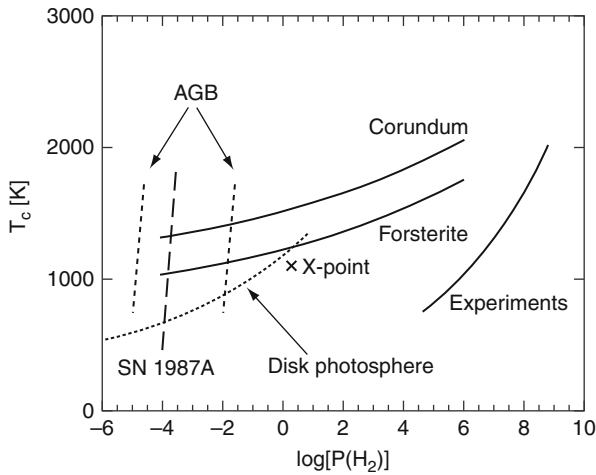
the first molecules and the subsequent growth on these condensation nuclei (Cherchneff and Dwek 2010). These studies promise a better understanding of the chemical growth of dust in these environments but they do require a critical evaluation against future (JWST) observations. In [Fig. 10-13](#), it is assumed that all silicon, iron, and carbon ends up in dust in supernovae (Tielens 2001).

Young stellar objects may be another prominent source of dust in the interstellar medium (Dwek and Scalo 1980). A fraction of the material that enters the protoplanetary disk is ejected through a jet/wind once it has been transported to a region close to the protostar. Because much of the mass loss may be in the form of a neutral wind which is difficult to detect, estimated values for the fraction ejected vary substantially. Models for the X-wind – as an example of magneto-centrifugally driven flows – suggest that one-third of the accreted material is ejected this way (Shu et al. 1994, and two-third is accreted into the central stellar object). Other models provide more conservative numbers of perhaps 0.1 for this fraction (Konigl and Pudritz 2000). In [Fig. 10-13](#), the higher value is adopted, providing an upper limit on the contribution by YSOs. Observationally, it is unclear whether these winds are dusty; let alone whether such dust is newly condensed rather than entrained from the surrounding natal cloud. However, the physical conditions and timescales associated with the inner region of protoplanetary disks are very conducive to dust formation ([Sect. 5.2](#)). Moreover, for the Solar Nebula, it is known that a large fraction of the material in the inner 2 AU has been extensively processed (e.g., vaporized and recondensed) and likely this is a general aspect of protoplanetary disks. [Figure 10-13](#) assumes that all the Si, Mg, and Fe in these winds is in the form of silicates.

5.2 Formation in Stellar Ejecta

Theoretical and experimental studies on the condensation and growth of dust are compared to astronomical observations in [Fig. 10-14](#). Traditionally, theoretical studies of dust formation and the composition of circumstellar dust are focused on the thermodynamic condensation sequence. Taking into account a wide variety of gas phase and solid state compounds, such

studies calculate the equilibrium composition of a gas of a given (Solar or non-Solar) elemental composition based upon measured and calculated thermodynamic properties of these materials (Grossman and Larimer 1974; Salpeter 1977). The results of such calculations are generally summarized using the condensation temperature – the temperature at which 50% of an element has condensed out – at a given pressure (● Fig. 10-14). These studies show that, for an O-rich gas, oxides (Al_2O_3 , corundum) are among the first condensates forming around 1,200 K at low pressures and at $\sim 2,000$ K at high pressures. From abundance considerations, the main condensate, magnesium-rich olivine (Mg_2SiO_4 , forsterite) forms at temperatures which are



■ Fig. 10-14

Theoretical, experimental, and astronomical studies on the condensation of silicates and oxides in a gas with Solar elemental composition. Total hydrogen pressures are in dynes cm^{-2} . The *solid curves* labeled corundum and forsterite represent the results of thermodynamic equilibrium calculations for the condensation temperature at a given pressure (Salpeter 1977). The results of laboratory studies on the critical pressure at which nucleation (of Si_2O_3) was detected in a SiO gas at a given temperature is shown as the *solid line* labeled experiments (Nuth and Donn 1982). These critical SiO pressures have been converted into H_2 pressures using the Si elemental abundance. The two almost vertical *short-dashed lines* labeled AGB show the relation between pressure and temperature in the circumstellar winds of Asymptotic Giant Branch stars around the sonic point for two different mass loss rates ($10^{-7} - 10^{-4} M_{\odot} \text{ year}^{-1}$). For these dust-driven winds, the bulk of the dust should condense close to the sonic point. The *dashed line* indicates the estimated gas pressure and temperature at day 615 in SN 1987A when dust formation was clearly proceeding (Wooden et al. 2008). The *top part* refers to the gas temperature in the C–O zone cooled by gaseous CO ($\sim 1,800$ K), while the *bottom part* is appropriate for the observed dust temperature (400 K). Calculated temperatures and pressures in the photosphere of a passively heated disk at a distance of 0.1–1 AU around a T-Tauri star are shown as a *dotted line* (Chiang and Goldreich 1999). For such an externally irradiated disk, temperatures in the interior are much less. Actively accreting disks will have much higher temperatures in their interiors (as well as in their photospheres). The X indicates the estimated conditions at the X-point in magnetocentrifugally driven flows from T-Tauri stars

some 200 K lower. These thermochemical models have been very successful in explaining the mineral composition of meteorites in the Solar system.

A variety of laboratory studies on the condensation of dust have been performed, mainly focusing on the physical properties of the condensed grains (Demyk et al. 2001; Fabian et al. 2000; Hallenbeck et al. 1998; Nuth and Donn 1982; Nuth et al. 2002; Rotundi et al. 2002). The results for the critical SiO pressure at which nucleation takes place in an SiO-H₂ gas (Nuth and Donn 1982) are compared to the thermodynamic equilibrium calculations in [Fig. 10-14](#). These critical pressures are displayed as equivalent H₂ pressures adopting the Solar elemental Si abundance. This assumes implicitly that H₂ does not actively participate in (the rate-limiting step of) the nucleation process, which is borne out by experiments. These experimental pressures for the onset of nucleation exceed the thermodynamic equilibrium calculations by many orders of magnitude. It is clear that non-thermodynamic considerations must play a key role in these experiments. Likely, at the short timescale for nucleation in these experiments (~10 s), the kinetics associated with the chemical pathways toward the critical nucleation clusters dominates. Based upon the extensive literature for soot chemistry in terrestrial environments, largely derived from the automotive and fuel industries, the chemical pathways toward carbon dust have been modeled in an astrophysical setting, using detailed reaction networks (Cherchneff et al. 1992, 2000; Frenklach and Feigelson 1989). In recent years, a comparable study of condensation in O-rich environments has been started because of its interest in stardust studies, supernova dust condensates – particularly in the early Universe – and dust formation in exoplanetary atmospheres (Cherchneff and Dwek 2010). These studies bear out the importance of kinetic considerations.

Astronomical observations show that silicate and oxide dust form readily in the ejecta of Asymptotic Giant Branch stars for mass loss rates in the range 10^{-7} – 10^{-4} M_⊙ year⁻¹. The physical conditions around the sonic point – where the bulk of the dust must condense – in these winds can be estimated from mass and momentum considerations ([Fig. 10-14](#)). At the lower mass loss rates ($\approx 10^{-7}$ M_⊙ year⁻¹), the newly formed dust is dominated by oxides such as Al₂O₃ and MgFeO, but for mass loss rates exceeding some 10^{-6} M_⊙ year⁻¹ amorphous silicates dominate the dust budget (Cami 2001; Sloan et al. 1998). The IR spectra of AGB stars with the highest mass loss rates ($\sim 10^{-4}$ M_⊙ year⁻¹) show evidence for crystalline forsterite and enstatite grains. It seems thus that, as the mass loss rate increases, the calculated thermodynamic condensation sequence is “followed” to lower temperatures, possibly reflecting the importance of freeze-out in the rapidly expanding AGB shells for lower mass loss rates (Cami 2001; Tielens et al. 1997). The astronomical record on dust formation in supernova ejecta is very limited. Dust nucleation and growth has only been observed for SN 1987A where it occurred starting at day 600 and proceeding through day ~800 (Wooden et al. 2008). The conditions in the chemically different zones of the ejecta were very different but likely this dust was formed in clumps in the Fe-rich and/or O-rich zones. The pressure in this zone has been calculated from a few simple considerations (Wooden et al. 2008). The temperature ranges from the observed dust temperature to the estimated gas temperature ($\approx 1,800$ K). Overall, the physical conditions where dust condensation and growth is observed in astronomical settings is in good agreement with those expected from thermodynamic equilibrium calculations and occurs at pressures which are much lower than the experimental critical pressures ([Fig. 10-14](#)). Possibly, this reflects the much longer timescales for dust formation in these astronomical settings – ranging from weeks for supernovae to years for AGB stars – than in the experiments. As a corollary, the experiments may have only limited value in terms of characterizing condensation products.

5.3 Processing by Interstellar Shocks

Interstellar shocks are an important destruction agent of interstellar dust (Jones et al. 1994, 1996) due to sputtering by impinging energetic ions. Supernova explosions drive strong shock waves into the surrounding interstellar medium. As the supernova remnant expands, the shock velocity drops. Initially, this expansion is adiabatic (the Sedov–Taylor phase) because very hot gas cools slowly, but when the shock velocity drops below $\approx 250 \text{ km s}^{-1}$, cooling becomes important (the radiative phase). Eventually, the remnant will merge with the ISM. Most of the destruction in the ISM is done by radiative shocks basically because a much larger volume is processed by low-velocity shocks than by fast shocks. The discussion here is focused on such shocks.

While the gas is stopped in the shock front, because of their inertia, dust grains will keep moving at three-fourth of the shock speed relative to the gas. Since the grains are charged (► Sect. 4.2), they will gyrate around the magnetic field. The Larmor radius, R_L , is approximately given by

$$R_L = \frac{m_d v_d c}{Z_d e B} \approx \frac{1.3 \times 10^{16}}{\phi_d / kT} \left(\frac{a}{1,000 \text{ \AA}} \right)^2 \left(\frac{v}{100 \text{ km s}^{-1}} \right) \left(\frac{10^5 \text{ K}}{T} \right) \text{ cm} \quad (10.60)$$

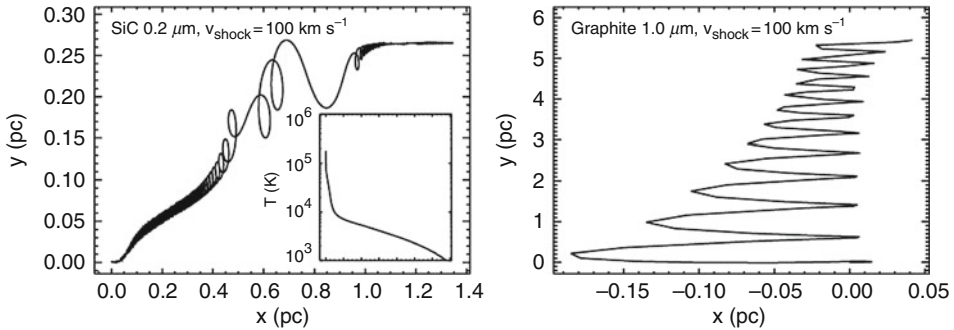
with m_d , Z_d , and v_d the grain mass, charge, and velocity and B the strength of the magnetic field, and where the grain potential, ϕ , (► 10.59) is typically of the same order as the thermal energy if collisions dominate and is ~ 60 if photoelectric effect plays a role. The size of the region of hot postshock gas is approximately, $2 \times 10^{17} (v_s/100 \text{ km s}^{-1})^{4.2} \text{ n cm}$. Thus, for typical interstellar grain sizes ($a < 0.3 \mu\text{m}$), grains are position coupled to the gas behind the shock. The magnetic moment, μ_d , of the grains,

$$\mu_d = \frac{m_d v_d^2}{2B} \quad (10.61)$$

is conserved and thus, as the gas compresses behind the shock because of cooling and the magnetic field strength increases because of flux freezing ($B \sim \rho$), the grains will spin up. This process, called betatron acceleration, is counteracted by gas drag. This drag is more effective for small grains which are therefore less susceptible to betatron acceleration. In all, grains will move at considerable speeds relative to the gas and to each other over much of the shock structure.

The trajectory of a typical ($< 0.3 \mu\text{m}$) interstellar grain in an interstellar shock is illustrated in ► Fig. 10-15. The Larmor radius scales with v_d/Z_d and slowly increases in the postshock gas. Downstream, a sudden charge reversal causes a change in spinning direction but by that time the grain has largely stopped. Smaller interstellar grains have less inertia and hence stop more quickly. All these grains are position coupled to the gas. This is different for large grains ($a > 1 \mu\text{m}$) where the Larmor radius can become comparable to the shock size. At that point, the grain can be reflected many times back and forth across the shock front, and every time it is accelerated to higher velocities (► Fig. 10-15). This Fermi acceleration process leads to very high velocities ($\sim 3,000 \text{ km s}^{-1}$) before the grain is completely sputtered away. Parenthetically, this process may be the first step in the production of cosmic rays since the resulting fast moving ions can be further Fermi-accelerated to cosmic ray energies. It has been suggested that the non-solar composition of cosmic rays may reflect this process (Ellison et al. 1997). For very big grains ($a > 10 \mu\text{m}$), the Larmor radius is so large that they traverse the shock without (dynamically) noticing its presence.

The relative motions between grains and gas in the postshock will sputter atoms from the grain. At grain–gas velocities of 100 km s^{-1} , impinging H-atoms have some 50 eV of energy and He atoms some 200 eV. At these energies, this will, typically, result in sputtering yields of about



■ Fig. 10-15

Calculated trajectories for individual dust grains in interstellar shocks (Slavin et al. 2004). These calculations are for plane-parallel shocks where the material enters from the *left* and the magnetic field is directed along the *z*-axis. The insert shows the calculated gas temperature as a function of the distance behind the shock front. The *x*-scale is the same as that in the trajectory panel. *Left panel:* A typical interstellar grain ($a = 0.2 \mu\text{m}$) is position coupled in a 100 km s^{-1} shock. *Right panel:* a $1 \mu\text{m}$ -sized grain is rapidly Fermi accelerated across the shock front and reaches velocities exceeding $1,000 \text{ km s}^{-1}$ before complete destruction. The atoms injected into the gas phase will be further accelerated by the shock and are likely the origin of galactic cosmic rays. Note the difference in *x*- and *y*-scale

0.01 per impinging H-atom. Ignoring, for simplicity, betatron acceleration and Coulomb drag, a grain will have to encounter its own mass in gas atoms in order to slow down. The fraction, f_{sput} , of a grain sputtered in the shock is then

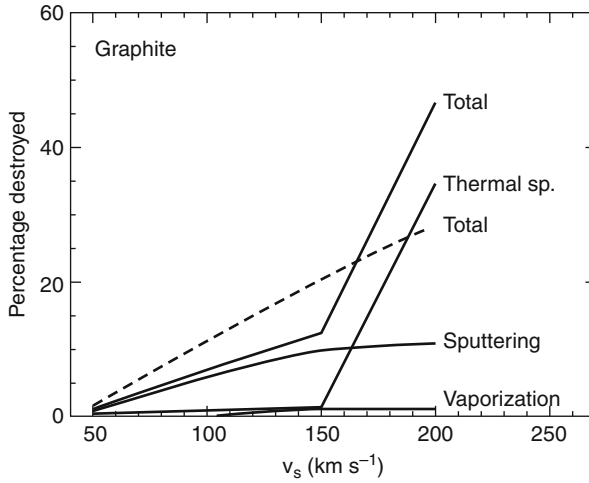
$$f_{\text{sput}} \simeq \bar{Y}_{\text{sput}} \frac{m_a}{\mu}, \quad (10.62)$$

with \bar{Y}_{sput} the average sputtering yield, μ the average atomic weight of gas atoms, and m_a the average atomic weight of an atom in the grain. So, a shock will sputter a layer with a thickness, Δa

$$\frac{\Delta a}{a} = \frac{1}{3} \bar{Y} \frac{m_a}{\mu} \simeq 0.03, \quad (10.63)$$

and the numerical value is appropriate for carbonaceous grains. Of course, betatron acceleration and Coulomb drag cannot really be ignored. Moreover, grain velocities are well above the threshold for cratering, melting, and vaporization in collisions between two grains and these processes have to be evaluated as well.

Over the years, a number of theoretical studies have appeared that detail the destructive effects of interstellar shocks (Jones et al. 1994, 1996). These studies calculate the shock structure, grain velocities, and sputtering rate and grain–grain collision rate as a function of position behind a shock of a given velocity. For a given grain size (distribution), this allows then the calculation of the fraction of a grain destroyed for shocks of different velocities. Representative results are shown in ► Fig. 10-16. Convolving these results with models for the frequency of shocks of different velocities, results then in the lifetime against destruction of interstellar dust grains. The results of these studies show that sputtering is the dominant process returning solid material to the gas phase. This is a slow “chipping” away of large grains where a 100 km s^{-1} shock will typically return some 10% of the grain mass to the gas phase. Small grains ($a < 100 \text{ \AA}$) are



■ Fig. 10-16

Graphite grain processing as a function of shock velocity. Separate curves for inertial sputtering due to relative gas-grain motions, thermal sputtering (due to impacting hot gas), and vaporization in grain-grain collisions are also shown (Figure reproduced from Jones et al. 1996)

hardly affected by such relatively frequent, low-velocity shocks (they are dragged to a halt before betatron acceleration kicks in) and are mainly (completely) destroyed by much less frequent, high-velocity ($v_s > 300 \text{ km s}^{-1}$) shocks. Grain-grain collisions are important in redistributing the grain mass from large grains to small grains. Indeed, a single 100 km s^{-1} shock will lead to the complete disruption of essentially all large ($300\text{--}3,000 \text{ \AA}$) grains considered (Jones et al. 1996). In contrast, vaporization by grain-grain collisions is of little consequence.

The destruction rate, k_{des} , of interstellar dust by supernova shock waves is given by

$$k_{\text{des}} M_{\text{ISM}} = \frac{1}{\tau_{\text{SN}}} \int \varepsilon(v_s) dM_s(v_s), \quad (10.64)$$

where M_{ISM} is the total mass of the ISM ($4.5 \times 10^9 M_{\odot}$), τ_{SN} is the effective interval between supernova explosions, $\varepsilon(v_s)$ is the fraction of dust destroyed by a shock of velocity v_s , and $M_s(v_s)$ is the mass shocked to a velocity of at least v_s . The mass processed by a SNR depends on the expansion of a supernova remnant in the interstellar medium. As an example, let us presume that dust destruction is dominated by 100 km s^{-1} shocks which destroy 10% of a grain. Such supernovae effectively occur once every 100 years in the Milky Way and process about $7 \times 10^3 M_{\odot}$ at velocities exceeding 100 km s^{-1} in a two-phase medium (Tielens 2005). This results then in a cumulative destruction timescale of $\sim 650 \text{ M year}$ for interstellar dust. Numerical evaluation of the dust destruction in supernova shocks coupled with models for the expansion of supernova remnants into the ISM come to very much the same result, $\approx 500 \text{ M year}$. This timescale is much less than the timescale (~ 2 billion years) at which grains are replenished by stellar sources (Dwek and Scalo 1980; Jones et al. 1994, 1996). Cumulative is a key word here since this assumes that this is the only process working and the effects of successive shocks can be added. This is, however, not the case in a multiphase interstellar medium.

Consider a two-phase medium consisting of clouds embedded in a low density intercloud medium which fills most of the volume. Supernova remnants mainly shock material in the intercloud medium to high velocities – shock velocities in clouds which are engulfed by a supernova remnant will be reduced by a factor $(\rho_{ic}/\rho_c)^{1/2} \simeq 0.1$ – and dust destruction is limited to the intercloud phase. This destruction is counteracted by accretion in the much denser cloud phase. Now, material is rapidly exchanged between these phases through evaporation of clouds, thermal instabilities, and/or turbulent compression/shearing. Timescales for interchange between the cloud and intercloud phase are estimated to be of order 30 M year. In steady state, the reverse interchange timescale is smaller by the ratio of the masses of these phases, $\simeq 0.1$ in the plane of the galaxy. Much of this exchange may also occur through the lower halo but timescales will still be short compared to the overall life cycle of dust (and gas) in the galaxy. Typically, for the Milky Way, in every sojourn into the intercloud medium, a parcel of gas is shocked by a 100 km s^{-1} shock (and about four 50 km s^{-1} shocks). Upon return to the cloud medium, the sputtered gas atoms can be reaccreted forming a thin protective “coating” which can be sputtered off again in the next cycle. Apart from this thin coating, the destruction timescale for grains can then be much larger than the $\simeq 500$ M year estimated from the cumulative effects of interstellar shocks.

5.4 Depletions and the Life Cycle of Dust

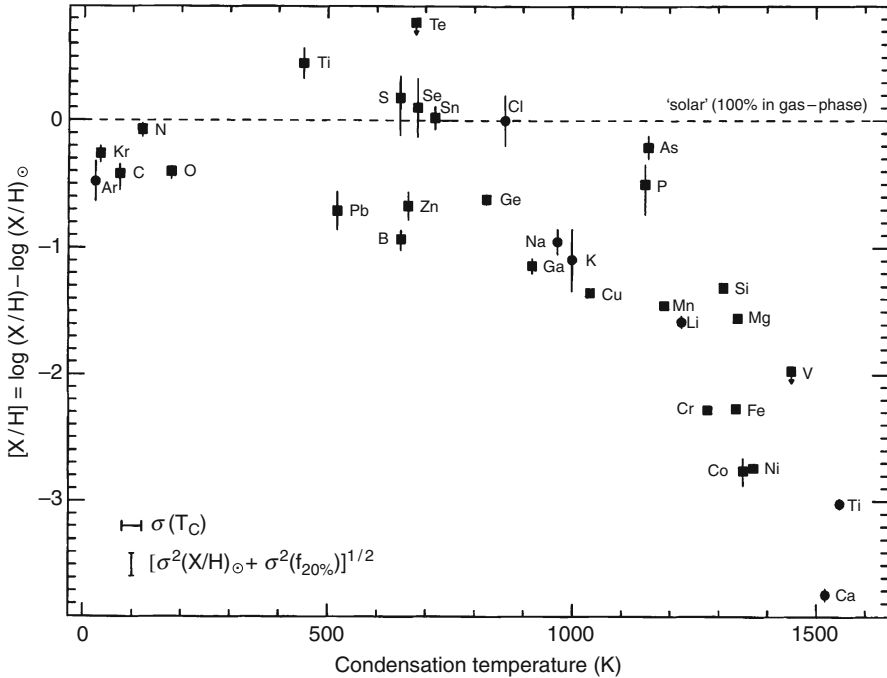
Measured elemental abundances and their variations provide insight in the characteristics of interstellar dust on the scale of individual clouds/interclouds and its evolution in the ISM. The data (► [Fig. 10-17](#)) does reveal a rough correlation between the depletion of an element and its condensation temperature. This is sometimes taken as an indication that depletions and dust composition are governed by formation processes in stellar ejecta. However, other processes in the dust life cycle will also roughly scale with the binding energy of an element to solids (e.g., sputtering) and this is not conclusive. Some elements are observed to be depleted by large factors; Titanium, calcium, and aluminum by some three orders of magnitude! Whatever process regulates their abundance must be very efficient. In this respect, some 10% of all matter in the ISM is injected by fast winds from O and B stars that do not form dust (Jura 1987). Moreover, supernova are also likely not very efficient dust producers (Barlow 2009). So, a substantial fraction of the elements are injected in gaseous form and are then depleted through accretion processes in the ISM itself and should be part of a thin outer coating (cf. ► [Sect. 5.3](#)).

Observations have also revealed a large and systematic difference in the elemental depletions in the different phases of the interstellar medium (► [Fig. 10-18](#); Cartledge et al. 2006; Savage and Sembach 1996) and these variations can be used to get an “observational” handle on the rate at which dust is destroyed in the intercloud medium and reformed through accretion in the cloud phase. Assume that the intercloud phase has a depletion, δ_i , (fraction of an element locked up in dust) and dust is destroyed with a rate, k_{des} . The cloud phase has a depletion, δ_c , and gas accretes onto grains at rate, k_{acc} . The two phases mix at rates k_1 ($i \rightarrow c$) and k_2 ($c \rightarrow i$). Dust injection occurs at a rate $\delta_o k_{in}$. Making a few simplifying assumption leads then to the following two equations,

$$\frac{\delta_c}{\delta_i} = \left(1 + \frac{k_{des}}{k_1} \left(\frac{\frac{1}{2} + k_{acc}/\delta_o k_{in}}{1 + k_{acc}/\delta_o k_{in}} \right) \right), \quad (10.65)$$

and

$$\frac{\delta_c}{1 - \delta_c} = \frac{k_{acc}}{k_2} \left(1 + \frac{k_1}{k_{des}} \right). \quad (10.66)$$



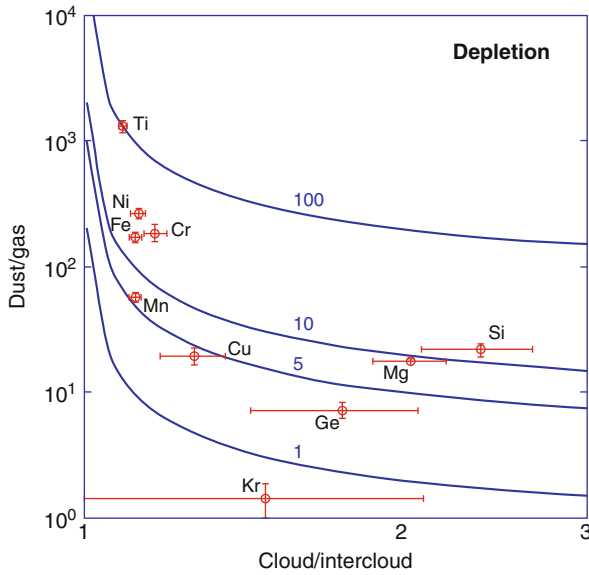
■ Fig. 10-17

Interstellar elemental depletions measured for a variety of species are shown as a function of the condensation temperature (i.e., the temperature at which 50% of this element condenses out in solid form in a cooling gas) (Figure adapted from Savage and Sembach 1996)

These two equations link relative and absolute cloud and intercloud depletions to ratios of rates for the various processes involved: destruction, accretion, injection, and mixing.

This simple model is compared to the observations in [Fig. 10-18](#). The observed large depletion variations demonstrate rather directly that the processes involved – shock destruction and accretion – operate on a timescale similar to the timescale at which material is mixed from the cloud to the intercloud medium and back (cf. [10.65] and [10.66]). This mixing timescale is much less (≈ 30 million years) than the timescale at which new dust is injected into the ISM. Thus, specifically, some 10% of the iron, 15% of the magnesium, and some 30% of the silicon is returned to the gas phase upon each sojourn into the intercloud medium and then rapidly reaccreted once the material is cycled back to the (diffuse) cloud phase. Again, as for the material injected in to the ISM by non-dust-producing stellar sources, this points toward the importance of the formation of a thin, outer layer. This thin coating is readily sputtered in the warm intercloud medium and in that way protects underlying (stardust) grains against the destructive effects of shocks.

Some further insights in these processes can be gleaned from this data. The ratio of destruction to the accretion rate varies along the model curves shown in [Fig. 10-18](#). At the low temperatures of the ISM, sticking coefficients of all elements (except for H and He) will be close to unity. The variations in the behavior of the different elements (e.g., Ti, Fe vs. Si and Mg) seem, therefore, to be driven by differences in destruction. Some of that may reflect the presence



■ Fig. 10-18

Observed depletions of elements in the ISM (Jenkins 2009; Savage and Sembach 1996). The y-axis is the ratio of the abundance of an element in the dust phase to its abundance in the gas phase, both measured in diffuse clouds. The x-axis is the ratio of the depletion of these elements in the diffuse cloud medium to that in the intercloud medium. The *solid lines* are the results of a simple model balancing destruction in the intercloud medium with accretion in the cloud medium (Tielens 1998). The labels indicate the adopted values for the accretion rate relative to the cloud-to-intercloud mixing ratio. As these observations demonstrate, the rates for destruction in the intercloud phase, accretion in the cloud phase, and mixing between these phases have to be within a factor of a few of each other and destruction and accretion are thus very rapid compared to the injection rate of dust by stars into the ISM

of multiple grain components which are sputtered at different rates in shocks. Alternatively, some of these elements may be preferentially locked up in cores (e.g., rutile (TiO_2), spinel (MgAl_2O_4), or corundum (Al_2O_3)) deeply buried inside, for example, magnesium silicates. Such layered grain structures have often been speculated to be the direct result of the condensation sequence in stellar ejecta where the most refractory materials form first and other compounds may heterogeneously condense out on top. Indeed, individual graphitic stardust grains have revealed direct examples of such structures (Bernatowicz et al. 2003). Nevertheless, recall that at least 10% of the atoms of all elements are locked up in a thin coating. The difference in abundance variations between cloud and intercloud phases for these elements must therefore reflect variations in destruction behavior; most likely, the effects of non-stoichiometric sputtering (Demyk et al. 2004). Finally, the depletion pattern of oxygen indicates that it participates in this shock-sputtering and reaccretion cycle. However, carbon does not show a difference in depletion between the cloud and intercloud medium. So, likely this thin outer layer has an oxide rather than a carbide structure. This difference in chemical behavior is not understood:

perhaps, accreted carbon is rapidly cycled to volatile compounds (e.g., CH₄, CO) that are readily photodesorbed rather than become integrated into a silicate or oxide network.

6 The Role of PAHs and Dust in the ISM

6.1 Photoelectric Heating of Interstellar Gas

Photoelectric heating is the dominant process that couples the energy balance of the gas to the nonionizing radiation field of stars in HI regions. As such, photo-electric heating ultimately controls the phase structure of the ISM and the physical conditions in PDRs – which includes most of the HI gas mass of the ISM – and therefore the evolution of the ISM of galaxies (Hollenbach and Tielens 1999). It has long been recognized that photoelectric heating is dominated by the smallest grains present in the ISM and PAHs and very small grains dominate the heating of interstellar gas (Bakes and Tielens 1994). Essentially, absorption of an FUV photon creates an electronhole pair in the material. The electron diffuses toward the surface, losing its excess kinetic energy along the way through “collisions.” Because the FUV absorption depth ($\sim 100 \text{ \AA}$) can be much larger than the mean free path of low energy electrons in solid materials, the yield is very low for large grains (cf. [Sect. 4.2.2](#)). Because of the Coulomb attraction, the photoelectric heating efficiency is sensitive to the grain charge (e.g., ionization potential). For simplicity, consider a species with two ionization stages, neutral and singly ionized. The photoelectric efficiency is then given by

$$\varepsilon = f(Z=0) \left(\frac{h\nu - IP}{h\nu} \right) = \left(\frac{1}{1 + 3.5 \times 10^{-6} N_c^{1/2} \gamma} \right) \left(\frac{h\nu - IP}{h\nu} \right), \quad (10.67)$$

where $f(Z=0)$ is the neutral fraction and the charge of a species is governed by the ratio of the ionization rate over the recombination rate which is proportional to the charging parameter, $\gamma = G_o T^{1/2} / n_e$ ([Sect. 4.2.3](#)).

Extensive theoretical calculations on the heating by an interstellar grain size distribution of PAHs and small grains, including the effects of charge, have been performed by Bakes and Tielens (1994) and the resulting efficiencies (ratio of gas heating to FUV absorption rate of grains and PAHs) have been fitted to a simple analytical formula

$$\varepsilon = \frac{4.87 \times 10^{-2}}{1 + 4 \times 10^{-3} \gamma^{0.73}} + \frac{3.65 \times 10^{-2} (T/10^4)^{0.7}}{1 + 2 \times 10^{-4} \gamma}. \quad (10.68)$$

The first term takes the ionization balance into account and is the equivalent of ([Sect. 10.67](#)). The dependence on γ is slightly less steep because larger PAHs have more charge states available. The second term introduces an additional temperature dependence, which reflects an increase in the electron recombination rate at high temperatures and the resulting decreased grain charge. This term enhances the efficiency by a factor 1.7 at $T \sim 10^4 \text{ K}$.

In terms of the efficiency ([Sect. 10.68](#)), the total photoelectric heating rate is given by

$$n \Gamma_{pe} = 10^{-24} \varepsilon n G_o \quad \text{erg cm}^{-3} \text{ s}^{-1}. \quad (10.69)$$

For a single species, in the small γ limit, the photoelectric heating scales with $G_o n$ (e.g., predominantly neutral species and ionization scales with G_o), while in the large γ limit, the heating rate

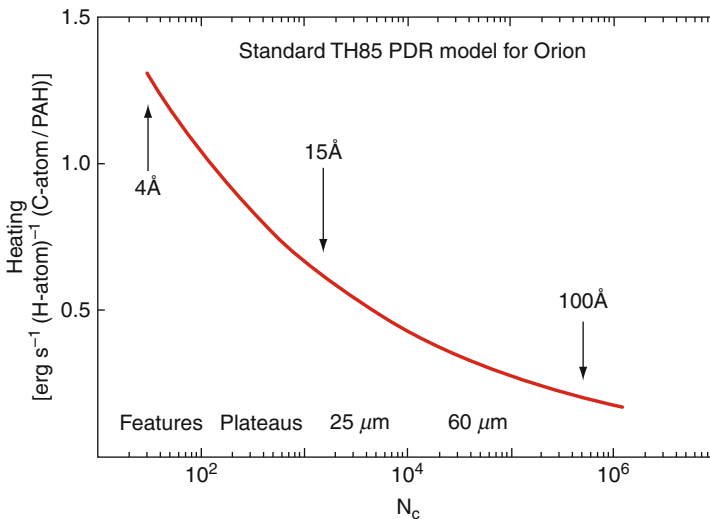
is independent of the intensity of the FUV field and proportional to nn_e through the recombination rate. For a size distribution, the latter is only approximately correct, because the transition from small to large γ occurs at different γ 's for different sized grains. For neutral PAHs and grains ($\gamma \ll 10^3 \text{ K}^{1/2} \text{ cm}^3$), a maximum efficiency of ~ 0.05 is reached and the photo-electric heating rate is then $\approx 5 \times 10^{-26} G_0 \text{ erg (H-atom)}^{-1} \text{ s}^{-1}$.

► *Figure 10-19* shows the calculated photoelectric heating rate as a function of grains size, illustrating that only species less than $\sim 100 \text{ \AA}$ contribute effectively to the photoelectric heating rate. ► *Figure 10-20* compares the calculated photoelectric efficiency as a function of the charging parameter, γ , to observations of the heating in diffuse sight lines and in well-known PDRs (Tielens 2009). There is some resemblance of the theory to the observations, providing some credence to the photoelectric heating model of interstellar gas.

6.2 Interstellar Molecules

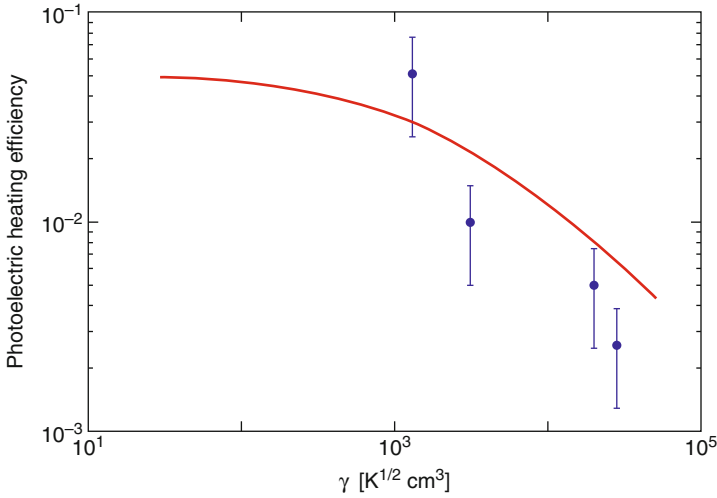
6.2.1 H₂ Formation

Molecular hydrogen is the most abundant molecule in space. It has been detected through its ultraviolet absorption lines and is very abundant in diffuse clouds for total columns in excess of



■ Fig. 10-19

The contribution to the photoelectric heating of interstellar gas by species of different sizes (Bakes and Tielens 1994), here traced by the number of carbon atoms, N_c . The results of these calculations are presented in such a way that equal areas under the curve correspond to equal contributions to the heating. Typically, about half of the heating originates from PAH and PAH clusters ($< 10^3$ C-atoms). The other half is contributed by very small grains ($15 < a < 100 \text{ \AA}$). Classical grains do not contribute noticeably to the heating. The typical IR emission characteristics as a function of size are indicated at the bottom of the figure



■ Fig. 10-20

The photoelectric efficiency as a function of the charging parameter, $\gamma = G_0 T^{1/2} / n_e$ (proportional to the ionization rate over the recombination rate). Neutral species are located to the *left* in this figure while to the *right* the charge of species increases. The data points indicate the measured heating efficiency for the diffuse ISM sight lines, ζ Oph and ξ Per, and the well-studied PDRs, NGC 2023, and the Orion Bar

$\sim 5 \times 10^{20}$ H-nuclei/cm². Molecular hydrogen has also been seen through its near infrared rovibrational transitions in gas that is exposed to nearby bright UV sources (IR fluorescence) or in warm gas associated with dense photodissociation regions or (molecular) shocks. Analyses of the UV lines arrive at a H₂ formation rate of $3 \times 10^{-17} n n_H \text{ cm}^{-3} \text{ s}^{-1}$.

Molecular hydrogen is difficult to form directly in a collision of two H-atoms, since the binding energy has to be radiated away and for a small (homonuclear) molecule, the probability that this occurs is very small. Alternative gas phase routes toward H₂ involve the H⁻ ion and do not face this problem (e.g., $\text{H} + \text{H}^- \rightarrow \text{H}_2 + \text{e}$) but then the abundance of this intermediate anion is very low. Because of the difficulty to form H₂ in the gas phase, it generally accepted that the formation of H₂ in the ISM proceeds on the surfaces of interstellar dust grains. The H₂ formation rate per unit volume, $R_d(\text{H}_2)$, can then be expressed as

$$R_d(\text{H}_2) = \frac{1}{2} S(T, T_d) \eta n_d \sigma_d n(\text{H}) v_H, \quad (10.70)$$

where $S(T, T_d)$ is the sticking probability of an H atom with temperature T colliding with a grain of temperature T_d , η is the probability that an adsorbed H atom will migrate over the grain surface, find another H atom, and form H₂ before evaporating from the grain surface, $n_d \sigma_d$ is the total grain surface area per unit volume, n_H is the H atom density, and $v_H = 1.5 \times 10^4 T^{1/2} \text{ cm s}^{-1}$ is the thermal speed of the H atoms. Typically, $n_d \sigma_d \approx 10^{-21} n \text{ cm}^{-2}$ and (● 10.70) becomes then,

$$R_d(\text{H}_2) \approx 5 \times 10^{-17} \left(\frac{T}{100 \text{ K}} \right)^{1/2} S(T, T_d) \eta(T_d) n n_H \quad \text{cm}^{-3} \text{ s}^{-1}, \quad (10.71)$$

Detailed evaluations of experimental studies on sticking of (noble) gas atoms on solid surfaces yield a sticking coefficient for atomic H in diffuse clouds ($T = 100$ K, $T_d = 15$ K) of $\simeq 0.6$. For lower gas temperatures, the sticking coefficient increases to $\simeq 1$. Based upon theoretical and experimental studies (Hollenbach and Salpeter 1971; Pirronello et al. 1999), the reaction probability, η , has been evaluated to be effectively unity for interstellar conditions in diffuse clouds (e.g., T in the range 5–20 K). H_2 formation on interstellar grain surfaces can thus explain quantitatively the observed H_2 abundances. However, at higher temperatures, physisorbed atomic hydrogen may evaporate before reaction can occur and η will then critically depend on the presence and characteristics of sites that can bind atoms more strongly (e.g., kinks, steps in the surface or chemisorbed sites, Cazaux and Tielens 2004; Hollenbach and Salpeter 1971). Thus, while observations show that molecular hydrogen is readily made in PDRs where temperatures are $T \simeq 500$ K and $T_d \simeq 75$ K, the theoretical and experimental underpinning of this is incompletely understood. At very high temperatures, H_2 formation may involve H-atoms chemically bonded to aromatic structures such as PAHs or HAC grains (Habart et al. 2004).

6.2.2 Interstellar Ices

Inside dense clouds, gaseous species colliding with a grain will stick, possibly undergo chemical reactions, and form an ice mantle consisting of simple molecules. The rate at which species, i , freeze out on grains is given by

$$R_{\text{acc}} = n_d \sigma_d n_i v_i \simeq 1.6 \times 10^{-17} \left(\frac{T}{10 \text{ K}} \frac{16 \text{ amu}}{m_i} \right)^{1/2} n_o n_i, \quad (10.72)$$

where a sticking probability of unity has been assumed, $n_d \sigma_d$ is the total grain surface area per unit volume ($\simeq 10^{-17} n_o \text{ cm}^2$ with n_o the density of H-nuclei), n_i , v_i , and m_i are the density, the thermal speed, and mass of the species. Thus, the timescale for freeze-out is $\simeq 2$ M year at a density of 10^4 cm^{-3} . The (fastest) dynamical evolution timescale, the free-fall timescale, is about $(10^4 \text{ cm}^{-3}/n_o)^{1/2}$ M year and, hence, at density in excess of 10^4 cm^{-3} , freeze-out will be rapid with obvious consequences for the ionization balance and cooling of the gas.

On a per grain basis, the increase in mass, m_d , of a grain of size, a , is given by

$$\frac{dm_d(a)}{dt} = 4\pi a^2 \rho_s \frac{da}{dt} = \pi a^2 n_c v_i m_i, \quad (10.73)$$

with ρ_s the specific density of the grain material. Thus, the increase in grain size due to accretion of gas phase species on a grain is independent of the size of the grain and all grains grow the same size mantle. Because the interstellar grain size distribution is so strongly weighted toward small grains, $n(a) \propto a^{-3.5}$ (● Sect. 3.2), this implies that most of the ice volume is on the smallest grains. Adopting 3×10^{-4} as the abundance of the condensibles relative to hydrogen, it can be concluded that the increase in grain size due to the formation of an ice mantle is relatively modest, $\simeq 175 \text{ \AA}$. Assuming that accretion also occurs on PAH-like species, the size of the ice mantle decreases to about 90 \AA .

The composition of the ices will be controlled by the composition of the accreting gas and the chemical routes enabled by migrating species. While most of the hydrogen is in the form of H_2 – which has limited reactivity at the low temperatures ($T \simeq 10$ K) of the interstellar medium – cosmic rays maintain a low abundance of atomic hydrogen which is very reactive. Carbon is largely in the form of the very stable molecule, CO, but cosmic-ray-produced UV

photons will break out a small fraction of the carbon into atomic form. The oxygen not locked up in CO will be mainly in atomic form. Nitrogen is in the form of N_2 . This species is stable against cosmic-ray-produced UV photons but cosmic rays will also ionize helium atoms and ionized helium can break the N_2 bond. ● *Table 10-4* summarizes the expected composition of the accreting gas. Confirming early theoretical studies (Tielens and Hagen 1982), experimental studies on grain surface chemistry routes have in recent years demonstrated the importance of hydrogenation of species such as CO (Fuchs et al. 2009; Watanabe et al. 2004). However, oxidation can also be of importance. This is in good agreement with the observed composition of interstellar ice mantles which reveal high abundances of H_2O , CH_3OH , H_2CO , CO, CO_2 , NH_3 , and CH_4 .

As ● *Table 10-4* demonstrates, the observed composition of interstellar ices is very different from that in the gas phase, attesting to an active grain chemistry. Besides the hydrogenation and oxidation processes involved in grain surface chemistry, accreted ices can also subsequently be processed in various ways. Highly energetic (~ 100 MeV/nucleon) cosmic ray ions can penetrate these clouds and directly interact with these ices. These cosmic rays can also electronically excite gaseous H_2 , which then decay through emission in the Lyman Werner bands ($^1\Sigma_g \leftarrow ^1\Sigma_u$ and $^1\Sigma_g \leftarrow ^1\Pi_g$). Some of these ultraviolet photons can be absorbed by the ices initiating a rich photochemistry. Near the surface of the cloud, UV photons penetrating from nearby stars will also contribute to this photolysis process. Finally, the newly formed star can heat its environment, promoting the diffusion and reaction of radicals produced by

● **Table 10-4**

The composition of accreting gas

Species	Gas phase abundance ^a	Ice abundance ^b	Ice/gas
H_2	0.5	–	
He ^c	0.2	–	
H ^d	$2/n^{-1}$	–	
CO	8 (–5)	2.5 (–5)	0.3
O ^e	2.4 (–4)	–	
C ^f	8 (–7)	–	
N ₂ ^g	8 (–5)	?	
H_2CO	2 (–8)	–	
CH_3OH	2. (–9)	$\sim 2.$ (–6)	~ 1 (3)
OH	3 (–7)	–	
H_2O	<7 (–9)	1.0 (–4)	>1 (4)
NH_3	2 (–8)	8. (–6)	4 (2)
HCN	2 (–8)	?	

^aGas phase abundances relative to all hydrogen nuclei derived from observations of the molecular cloud, TMC-1, unless otherwise noted

^bIce abundances in quiescent clouds

^cCosmic abundance of helium

^dTheoretical estimate of atomic hydrogen abundance from balance of cosmic ray dissociation and grain surface formation

^eCosmic abundance of oxygen minus oxygen locked up in silicates and in gaseous CO

^fTheoretical estimate of atomic carbon abundance from balance of photodissociation by cosmic-ray-produced UV photons and CO formation

^gCosmic abundance of nitrogen

photolysis as well as polymerization reactions. The relative importance of these processes is not fully understood. Nevertheless, it is clear that grains promote a diverse organic inventory in the molecular Universe.

6.2.3 PAHs and Interstellar Molecules

Interstellar PAHs are often thought to be injected into the interstellar medium by cool C-rich stellar outflows associated with Asymptotic Giant Branch (AGB) stars (Latter 1991) as a key intermediary or by-product of the soot formation process in such environments (Cherchneff et al. 1992; Frenklach and Feigelson 1989). The PAH family injected into the ISM is likely to be wide and varied but subsequent processing by the omnipresent UV photons will rapidly weed out the less stable species (► Fig. 10-21). Indeed, the IR spectral characteristics of post-AGB objects are typically very different from the spectra observed from PAHs in the interstellar medium (Sloan et al. 2007) and this has been attributed to the effects of this chemical weeding process (Pino et al. 2008) in the ISM. Processing by energetic ions in strong shock waves driven by supernova explosions may also contribute greatly to funneling the injected PAH distribution into its “strongest” members (Micelotta et al. 2010a). Either process (UV photo absorption or ion bombardment) will leave the PAH in a highly vibrational excited state where it is prone to fragmentation. However, this excited species may also cool through the emission of IR vibrational photons (cf. ► Sect. 4.1.2). Several channels for fragmentation can be open depending on excitation energy; e.g., H-loss and C₂H₂ loss (Ekern et al. 1998). Ion bombardment generally transfers more energy into internal energy of the species than available through UV photons in the ISM (with energies <13.6 eV) and hence may well open up different fragmentation channels.

Schematically, the fragmentation process can be written as



where PAH-R* is the excited species which can stabilize through emission of IR photons or through fragmentation. R is a sidegroup (e.g., H, CH₃, OH) or even a C₂H₂ molecule. There are various ways to evaluate the unimolecular dissociation rate constant for this process. Here, the rate constant is written in Arrhenius form,

$$k(E) = k_o(T_e) \exp\left[\frac{-E_o}{kT_e}\right], \quad (10.75)$$

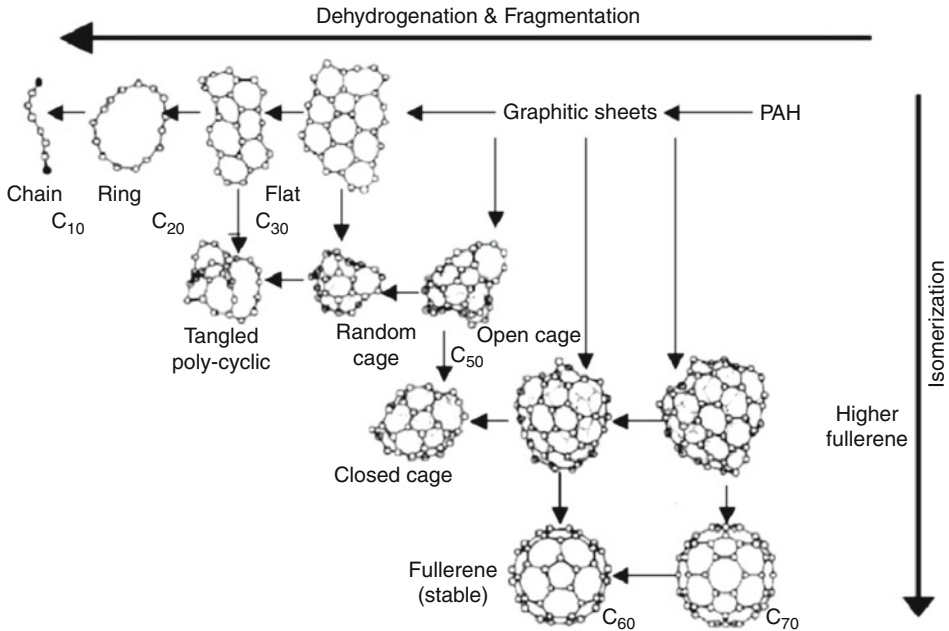
where T_e is an effective excitation temperature, E_o , the Arrhenius energy describing the process, and the preexponential factor, k_o , depends on the interaction potential (in the reverse reaction). Because the typical energy involved in these reactions is a fair fraction of the total energy in the system, a correction has to be made to the excitation temperature, T_m , describing the excitation of the vibrational modes (► 10.29). The finite heat bath correction results in

$$T_e = T_m \left(1 - 0.2 \frac{E_o}{E}\right). \quad (10.76)$$

The preexponential factor can be set equal to

$$k_o = \frac{kT_e}{h} \exp\left[1 + \frac{\Delta S}{R}\right], \quad (10.77)$$

with ΔS the entropy change. Typically, k_o is $\approx 3 \times 10^{16} \text{ s}^{-1}$. The energy parameters, E_o , cannot be easily evaluated from theoretical calculations but rather have to be determined by fitting experimental data. Unfortunately, experimental fragmentation studies have been limited to very small



■ Fig. 10-21

The chemical evolution of PAHs in the interstellar medium under the influence of UV photons and energetic ion bombardment combines the effects of dehydrogenation and fragmentation with those of isomerization. Fully hydrogenated PAHs – injected by stars into the ISM – are at the *top right* side. UV photolysis and ion bombardment will lead to complete H-loss and the formation of graphitic sheets. These sheets may isomerize to various types of cages or even fullerenes depending on internal energies. Further fragmentation may also lead to the formation of flats, rings, and chains

PAHs (<24 C-atoms) and the relevance of these studies to the larger PAH in space is unclear (and doubtful). Those studies do show, though, that E_o cannot be equated to the binding energy of the group under consideration. For small PAHs, H-loss is well described by E_o is 3 eV. The probability for dissociation depends then on the competition between fragmentation and IR photon emission,

$$p_d(E) = \frac{k(E)}{k(E) + k_{ir}(E)}, \quad (10.78)$$

The total fragmentation rate is then

$$k_{frag} = p_d(E) k_{uv}(E), \quad (10.79)$$

where $k_{uv}(E)$ is the absorption rate of UV photons with energy, E . This rate should then be integrated over the absorption spectrum of dissociating photons.

Adopting a guesstimate for E_o of 4.5 eV for acetylene loss (An E_o of 3.65 eV has been derived for small PAHs with open structures but that may not be very relevant for large compact PAHs.), $k_o = 3 \times 10^{16} \text{ s}^{-1}$, and $k_{uv} = 7 \times 10^{-10}$ per C-atom per second, k_{frag} is approximately given by $5 \times 10^{-8} \text{ year}^{-1}$ for a 50 C-atom PAH or a lifetime of some 20 M year against C-loss in the diffuse

ISM. H-loss for this PAH is much more rapid ($\approx 1 \text{ year}^{-1}$, for $E_o = 3.3 \text{ eV}$, Tielens 2005). Under interstellar conditions, fragmentation will preferentially occur through the “weakest link” and this is expected to lead to (almost) complete H-loss (► Fig. 10-21). Isomerization and further processing of these dehydrogenated PAHs will then favor the formation of fullerenes, cages, and/or rings, and chains depending on the internal energy. Some of these intermediary products may be very stable and become prominent members of the interstellar molecular inventory.

In this “trickle down” chemistry, interstellar PAHs may “father” a wide variety of related species but under the most extreme conditions, only the most stable species would dominate. This may be of particular relevance in regions of high UV field intensity – e.g., near the surfaces of photodissociation regions (PDRs) – or near supernova remnants. Observationally, the brightest PDR surfaces are characterized by very similar and unique infrared emission spectra, possibly reflecting this weeding down of the interstellar PAH family to the most stable members and/or reaction products. In addition to these stable products, fragmentation of PAHs could be the source of the unexpectedly high abundance of small hydrocarbon radicals in these environments (Pety et al. 2005). Likewise, dehydrogenated PAHs have often been considered as key to understanding the spectral pattern in the infrared emission features in particular for the CH out-of-plane bending modes (Duley and Williams 1981; Pauzat et al. 1997) and the diffuse interstellar bands (DIBs) (Duley 2006). However, except for C_{60} itself (Sellgren et al. 2007), the contributions of fullerenes and carbon cages to the interstellar emission spectra have not been considered. In a large measure, this reflects the fact that their spectra have not been studied systematically.

Revisiting the life cycle of interstellar PAHs, the UV destruction timescale for PAHs is calculated to be $\sim 100 \text{ M year}$. Destruction by shocks and cosmic rays occurs on a similar timescale (Micelotta et al. 2010a, b). The uncertainty in E_o , which enters in all of these estimates similarly, should be kept in mind in evaluating this. Nevertheless, it seems inescapable that PAHs should be reformed rapidly in the ISM itself either through active chemistry – via chemical routes that are not clear – or through fragmentation in grain–grain collisions in interstellar shocks (Jones et al. 1996).

7 Conclusions and Outlook

Interstellar PAHs and dust are deeply interwoven into the fabric of the Universe. They are at the root of many of the complex interplaying processes that drive the evolution of the interstellar medium and thereby the evolution of galaxies. While much progress has been made in our understanding of PAHs and dust over the last decades – largely driven by ever better observational opportunities from space – many questions remain. These include:

- What are the important stellar sources of dust and PAHs? Specifically, how important are high mass stars (red supergiants, luminous blue variables, supernova) versus low mass stars (asymptotic giant branch stars)? Does mantle formation in the ISM lead to a thin layer of veneer or can this be a dominant source of dust?
- What processes play a role in the evolution of dust and PAHs in the interstellar medium?
- What are the characteristics of interstellar PAHs and dust? How does that depend on metallicity, star formation rate of the galaxy, and ISM conditions (e.g., density, UV field, cosmic ray flux, turbulence, . . .)?

- What kind of dust entered the Solar Nebula and other protoplanetary disks? What are the characteristics of PAHs in regions of planet formation, particularly in the habitable zone? What processes played a role in their evolution in the protoplanetary systems? Do the characteristics of the dust influence the formation of the resulting planetary bodies?
- What is the role of PAHs and very small grains in the energy and ionization balance of the ISM? And how does that influence the structure of the ISM and the star formation activity?
- How are dust and PAHs affected in extreme environments such as in AGN accretion disks, near GRB, or in starbursts?
- How did the characteristics of dust and PAHs evolve with time in the Universe and how did this evolution affect the evolution of galaxies, stars, and planets?

Over the next decade, a further deepening can be expected of our understanding of interstellar PAHs and dust and their role in the Universe. The launch of the Herschel Space Observatory has opened up the cool and dark Universe for systematic studies in the far-infrared and sub-millimeter. This will provide much better insight in the coldest dust content of interstellar and circumstellar media. The Stratospheric Observatory For Infrared Astronomy (SOFIA), in combination with Gaia, will be able to probe the inventory of dust injection by stellar sources in the Milky Way to much greater accuracy than hitherto possible. The James Webb Space Telescope can do this on the scale of individual galaxies in the local group. SOFIA is also well geared toward large-scale, spectral imaging studies of the distribution and characteristics of PAHs in space, while JWST can probe the changes in the PAH population on small scales such as in the inner planet forming zones around young stars. JWST can also for the first time probe a statistically meaningful sample of supernovae for the dust mass and determine their properties. In all, over the next decade, these new observing facilities will be able to answer many of the questions raised above.

Acknowledgements

It is a pleasure to thank the many students, postdocs, and colleagues that I have been privileged to work with over the last three decades. I particularly want to single out here Lou Allamandola, Jan Cami, Jean Chiar, Sacha Hony, Ant Jones, and Els Peeters for their untiring efforts to educate me in all things dusty and molecular in the Universe.

Cross-References

- [Astrophysics of Galactic Charged Cosmic Rays](#)
- [Galactic Distance Scales](#)
- [Galactic Neutral Hydrogen](#)
- [Gamma-Ray Emission of Supernova Remnants and the Origin of Galactic Cosmic Rays](#)
- [High-Velocity Clouds](#)
- [Magnetic Fields in Galaxies](#)
- [Mass Distribution and Rotation Curve in the Galaxy](#)

References

- Aannestad, P. A. 1989, in *Evolution of Interstellar Dust and Related Topics*, eds. A. Bonetti, J. M. Greenberg, S. Aiello (Amsterdam: Elsevier), 121
- Allamandola, L. J., & Tielens, A. G. G. M., eds. 1989, *Interstellar Dust* (Dordrecht: Reidel)
- Anders, E., & Zinner, E. 1993, *Meteoritics*, 28, 420
- Bakes, E. L. O., & Tielens, A. G. G. M. 1994, *ApJ*, 427, 822
- Bakes, E. L. O., Tielens, A. G. G. M., & Bauschlicher, C. W. 2001, *ApJ*, 556, 501
- Barlow, M. J. 2009, *Astrophysics in the Next Decade*, eds. H. Thronson, S. Massivo, & A. G. G. M. Tielens, *Astrophys Space Sci Proc.* (Netherlands: Springer Verlag), 247
- Bernatowicz, T. J., Messenger, S., Pravdivtseva, O., Swan, P., & Walker, R. M. 2003, *Geochim Cosmochim Acta*, 67, 4679
- Bohlin, R. C., Savage, B. D., & Drake, J. F. 1978, *ApJ*, 224, 132
- Bohren, C. F., & Huffman, D. R. 1983, *Absorption and Scattering of Light by Small Particles* (New York: Wiley)
- Boogert, A. C. A., & Ehrenfreund, P. 2004, *Astrophysics of Dust*, 309 (San Francisco: Astronomical Society of the Pacific), 547
- Boulanger, F. 2000, in *ESA-SP, 455, ISO Beyond the Sources: Studies of Extended Infrared Emission*, eds. R. J. Laureijs, K. Leech, & M. F. Kessler, 3
- Cami, J. 2001, PhD thesis, University of Amsterdam
- Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, *ApJ*, 345, 245
- Cartledge, S. I. B., Lauroesch, J. T., Meyer, D. M., & Sofia, U. J. 2006, *ApJ*, 641, 327
- Cazaux, S., & Tielens, A. G. G. M. 2004, *ApJ*, 604, 222
- Cherchneff, I., & Dwek, E. 2010, *ApJ*, 713, 1
- Cherchneff, I., Barker, J. R., & Tielens, A. G. G. M. 1992, *ApJ*, 401, 269
- Cherchneff, I., Le Teuff, Y. H., Williams, P. M., & Tielens, A. G. G. M. 2000, *A&A*, 357, 572
- Chiang, E. I., & Goldreich, P. 1999, *ApJ*, 519, 279
- Chiar, J. E., Tielens, A. G. G. M., Whittet, D. C. B., Schutte, W. A., Boogert, A. C. A., Lutz, D., van Dishoeck, E. F., & Bernstein, M. P. 2000, *ApJ*, 537, 749
- Clayton, G. C., Gordon, K. D., & Wolff, M. J. 2000, *ApJS*, 129, 147
- Compiegne, M., Flagey, N., Noriega-Crespo, A., Martin, P. G., Bernard, J.-P., Paladini, R., & Molinari, S. 2010, *ApJ*, 724, L44
- Demyk K., et al. 2001, *A&A*, 368, L38
- Demyk K., d'Hendecourt, L., Keroux, H., Jones, A. P., & Borg, J. 2004, *A&A*, 420, 547
- Désert, F. X., Boulanger, F., & Puget, J. L. 1990, *A&A*, 237, 215
- Draine, B. T. 2003, *Ann Rev Astron Astroph*, 41, 241
- Draine, B. T., & Lee, H. M. 1984, *ApJ*, 285, 89
- Draine, B. T., & Li, A. 2001, *ApJ*, 551, 807
- Draine, B. T., & Li, A. 2007, *ApJ*, 657, 810
- Draine, B. T., Sutin, B. 1987, *ApJ*, 320, 803
- Draine, B. T., et al. 2007, *ApJ*, 663, 866
- Duley, W. W. 2006, *ApJ*, 643, L21
- Duley, W. W., & Williams, D. A. 1981, *MNRAS*, 196, 269
- Dwek, E., & Scalo, J. M. 1980, *ApJ*, 239, 193
- Ekern, S.P., et al. 1998, *J Phys Chem A*, 102, 3498
- Ellison, D. C., Drury, L. O. C., & Meyer, J.-P. 1997, *ApJ*, 487, 197
- Fabian, D., Jäger, C., Hennning, Th., Dorschner, J., & Mutschke, H. 2000, *A&A*, 364, 282
- Frenklach, M., & Feigelson, E. D. 1989, *ApJ*, 341, 372
- Fuchs, G. W., Cuppen, H. M., Ioppolo, S., Romanzin, C., Bisschop, S. E., Andersson, S., van Dishoeck, E. F., & Linnartz, H. 2009, *A&A*, 505, 629
- Galliano, F., Madden, S. C., Jones, A. P., Wilson, C. D., & Bernard, J.-P. 2005, *A&A*, 434, 867
- Gordon, K. D., Clayton, G. C., Misselt, K. A., Landolt, A. U., & Wolff, M. J. 2003, *ApJ*, 594, 279
- Grossman, L., & Larimer, J. W. 1974, *Rev Geophys Space Phys*, 12, 71
- Habart, E., Boulanger, F., Verstraete, L., Walmsley, C. M., & Pineau des Forêts, G. 2004, *A&A*, 414, 531
- Hallenbeck, S. L., Nuth, J. A., Daukantas, P. L. 1998, *Icarus*, 131, 198
- Henkel, T., King, A., & Lyon, I. 2007, *Lunar and PIS Conf. Abstr.*, Vol. 38, (Houston: Lunar Planetary Institute), 2351
- Henning, T., Grün, E., & Steinacker, J. 2009, *cosmic dust: near and far*, *ASP Conf. Ser.*, Vol. 414 (San Francisco: Astronomical Society of the Pacific)
- Hollenbach, D., & Salpeter, E. E. 1971, *ApJ*, 163, 155
- Hollenbach, D. J., & Tielens, A. G. G. M. 1999, *Rev Mod Phys*, 71, 173
- Jenkins, E. B. 2009, *ApJ*, 700, 1299
- Jones A. P., Tielens, A. G. G. M., Hollenbach, D. J., & McKee, C. F. 1994, *ApJ*, 433, 797
- Jones, A. P., Tielens, A. G. G. M., & Hollenbach, D. J. 1996, *ApJ*, 469, 740
- Jura, M. 1987, in *ASSL, 134, Interstellar Processes*, eds. H. Thronson, & D. J. Hollenbach (Dordrecht: Kluwer), 3
- Kim, S.-H., Martin, P. G., & Henry, P. D. 1994, *ApJ*, 422, 164

- Konigl, A., & Pudritz, R. E. 2000, in *Protostars and Planets IV*, eds. V. Manning, A. P. Boss, & S. S. Russell (Tucson: University of Arizona Press), 759
- Kozasa, T., Hasegawa, H., & Seki, J. 1984 *Ap&SS*, 98, 61
- Krügel, E. 2002, *The Physics of Interstellar Dust* (Bristol: IOP)
- Latter, W. B. 1991, *ApJ*, 377, 187
- Li, A., & Draine, B. T. 2001, *ApJ*, 554, 778
- Mathis, J. S., Rumpl, W., & Nordsieck, K. H. 1977, *ApJ*, 217, 425
- Meikle, W. P. S., et al. 2007, *ApJ*, 665, 608
- Micelotta, E. R., Jones, A. P., & Tielens, A. G. G. M. 2010a, *A&A*, 510, A36
- Micelotta, E. R., Jones, A. P., & Tielens, A. G. G. M. 2010b, *A&A*, 510, A37
- Nittler, L. R., Alexander, C. M. O. D., Gao, X., Walker, R. M., & Zinner, E. 1997, *ApJ*, 483, 475
- Nuth, J. A., & Donn, B. 1982, *J Chem Phys*, 77, 2639
- Nuth, J. A., Rietmeijer, F. J. M., & Hill, H. G. M. 2002, *Meteoritics*, 37, 1579
- Pauzat, F., Talbi, D., & Ellinger, Y. 1997, *A&A*, 319, 318
- Peeters, E., Hony, S., Van Kerckhoven, C., Tielens, A. G. G. M., Allamandola, L. J., Hudgins, D. M., & Bauschlicher, C. W. 2002a, *A&A*, 390, 1089
- Peeters, E., et al. 2002b, *A&A*, 381, 571
- Pety, J., Teyssier, D., Fossé, D., Gerin, M., Roueff, E., Abergel, A., Habart, E., & Cernicharo, J. 2005, *A&A*, 435, 885
- Pino, T., et al. 2008, *A&A*, 490, 665
- Pirronello, V., Liu, C., Roser, J. E., & Vidali, G. 1999, *A&A*, 344, 681
- Rho, J., et al. 2008, *ApJ*, 673, 271
- Rho, J., Reach, W. T., Tappe, A., Hwang, U., Slavin, J. D., Kozasa, T., & Dunne, L. 2009, *ApJ*, 700, 579
- Rotundi, A., Brucato, J. R., Colangeli, L., Ferrini, G., Mennella, V., Palombo, E., & Palumbo, P. 2002, *Meteoritics*, 37, 1623
- Salpeter, E. E. 1977, *AnnRev Astron Astrophys*, 15, 267
- Savage, B. D., & Sembach, K. R. 1996, *Ann Rev Astron Astrophys*, 34, 279
- Schnaiter, M., Mutschke, H., Dorschner, J., Henning, Th., & Salama, F. 1998, *ApJ*, 498, 486
- Sellgren, K., Uchida, K. I., & Werner, M. W. 2007, *ApJ*, 659, 1338
- Shu, F., Najita, J., Ostriker, E., Wilkin, F., Ruden, S., & Lizano, S. 1994, *ApJ*, 429, 781
- Slavin, J. D., Jones, A. P., & Tielens, A. G. G. M. 2004, *ApJ*, 614, 796
- Sloan, G. C., & Price, S. D. 1998, *ApJS*, 119, 141
- Sloan, G. C., et al. 2007, *ApJ*, 664, 1144
- Steglich, M., Jäger, C., Rouillé, G., Huisken, F., Mutschke, H., & Henning, T. 2010, *ApJL*, 712, L16
- Sugerman, B. E. K., et al. 2006, *Science*, 313, 196
- Tielens, A. G. G. M. 1998, *ApJ*, 499, 267
- Tielens, A. G. G. M. 2001, in *Tetons 4: Galactic Structure, Stars and the Interstellar Medium*, Vol. 231, eds. C. E. Woodward, M. D. Bica, & J. M. Shull (San Francisco: ASP), 92
- Tielens, A. G. G. M. 2005, *The Physics and Chemistry of the Interstellar Medium* (Cambridge, UK: Cambridge University Press)
- Tielens, A. G. G. M. 2008, *Ann Rev Astron Astrophys*, 46, 289
- Tielens, A. G. G. M. 2009, *Astrophysics in the Next Decade*, eds: H. Thronson, S. Massimo, & A. G. G. M. Tielens, in *Astrophysics and Space Science Proceedings*, Vol. 271 (Netherlands: Springer Verlag)
- Tielens, A. G. G. M., & Hagen, W. 1982, *A&A*, 114, 245
- Tielens, A. G. G. M., Waters, L. B. F. M., Molster, F. J., & Justtanont, K. 1997, *ApSS*, 255, 415
- van de Hulst, H. C. 1957, *Light Scattering by Small Particles* (New York: Wiley)
- Watanabe, N., Nagaoka, A., Shiraki, T., & Kouchi, A. 2004, *ApJ*, 616, 638
- Whitter, D. C. B. 2003 *Dust in the Galactic Environment* (Bristol: IOP)
- Witt, A. N., Clayton, G. C., & Draine, B. T., eds. 2004, *Astrophysics of Dust* (San Francisco: Astronomical Society of the Pacific)
- Wooden, D. H., et al. 1993, *ApJS*, 88, 477
- Zinner, E. K. 2003, in *Treatise on Geochemistry*, Vol. 1, ed. K. K. Turekian, H. D. Holland, & A. M. Davis (Amsterdam: Elsevier), 17
- Zubko, V., Dwek, E., & Arendt, R. G. 2004, *ApJS*, 152, 211

11 Galactic Neutral Hydrogen

John M. Dickey

School of Mathematics and Physics, University of Tasmania,
Hobart, TAS, Australia

1	<i>Background</i>	550
1.1	Structures in the Atomic ISM	552
2	<i>Instruments and Techniques for 21-cm Line Observing</i>	553
3	<i>Radiative Transfer in the 21-cm Line</i>	555
3.1	Brightness Temperature	555
3.2	H I Cloud Masses	557
3.3	Optical Depth	558
3.4	H I Self-absorption	564
3.5	The Relationship Between H I Emission and Absorption	565
4	<i>The Longitude–Velocity Diagram and the Velocity Gradient</i>	566
4.1	Kinematics in a Circularly Rotating Disk	567
4.2	Rotation Curve Models	568
4.3	Modeling the Longitude–Velocity Diagram	569
5	<i>The Structure of the H I Disk</i>	571
5.1	The z Distribution in the Solar Neighborhood	571
5.2	The Outer Galaxy	574
6	<i>Thermal Equilibrium in the H I</i>	575
7	<i>Small-Scale Structure in the H I Medium</i>	580
7.1	The Spatial Power Spectrum	580
7.2	Tiny Scale Structure	581
8	<i>Looking Ahead</i>	583
	<i>References</i>	585

Abstract: The neutral atomic hydrogen in the Milky Way constitutes two different thermal phases of the interstellar gas: the warm neutral medium and the cool neutral medium. The best way to trace these phases on the scale of the entire Galactic disk is by using the $\lambda 21$ -cm line at radio frequencies. This chapter explains with examples how observations of the 21-cm line are interpreted, with emission and absorption spectra leading to estimates of the column density and the excitation- or spin-temperature of the gas.

The structure of the H I disk has both a thin component, with half-width ~ 100 pc that includes both warm and cool gas, plus a thick disk, ~ 400 pc half-width, that is mostly all warm. In the inner Galaxy the distribution of brightness in longitude-velocity coordinates is largely determined by the velocity gradient along the line of sight, with self-absorption by cool clouds competing with irregularities in the density and velocity field to modulate the intensity of the 21-cm emission. In the outer Galaxy the disk becomes more and more H I dominated, with an increase in the scale height and a strong warp that becomes very one-sided at Galactic radius greater than ~ 20 kpc.

The thermal balance in the two atomic phases is set by heating and cooling equilibrium, but the system does not have time to reach hydrostatic balance before supernova remnants disrupt the gas and broaden the pressure distribution function. Pressure variations can drive the gas either way between warm and cool phases, but the net flux must be from warm to cool, and on to molecular clouds, in order to explain the continuous cycle of star formation and chemical enrichment of the Galaxy.


The small-scale structure of the H I shows a power-law spectrum of velocity and density fluctuations similar to that of a turbulent fluid. In some cases, the structure is enhanced in the cool neutral medium on scales of a few hundred astronomical units. The magnetic field may be important dynamically, particularly on the smaller scales. These and other issues about the H I disk will be better answered when the next generation of telescopes becomes available for Galactic survey projects.

Keywords: 21-cm line, Cloud random velocities, Column density, Cool neutral medium, Cooling rate, Galactic plane surveys, Galactic structure, Galactic warp, Gas scale height, Heating rate, H I column density, HISA, Interstellar clouds, Interstellar magnetic field, Interstellar medium, Interstellar turbulence, Radio astronomy, Radio spectroscopy, Rotation curve, Spatial power spectrum, Spectral lines, Spin temperature, Radio interferometers, Square kilometer array, Thermal balance, Warm neutral medium

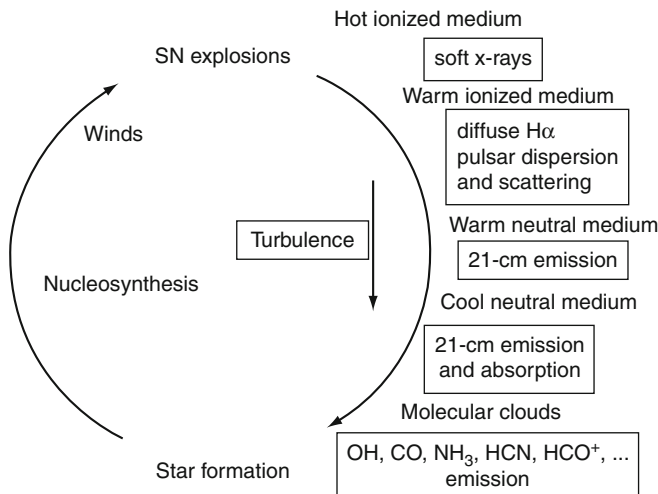
1 Background


Spiral galaxies look very different depending on the wavelength used to study them. In many spectral lines and in some continuum bands, galaxy images are dominated by the places where the most energy is produced: the massive stars and the regions where they form. At other wavelengths the light is a better tracer of the mass, at least of the baryonic mass that is mostly found in low mass stars. The 21-cm line of atomic hydrogen (H I) is not like either of these. Its brightness is biased *against* regions of massive star formation, and its large-scale distribution is quite different from that of any kind of stars. More than any other interstellar medium (ISM) tracer, the H I is widespread throughout spiral galaxy disks; in the outer Milky Way (MW) it extends

far beyond most of the stars and all other spectral lines that trace the gas. Structures large and small in the MW disk and halo can be traced in the H I with no corresponding features either in starlight or molecular line emission.

The H I is useful as an ISM tracer because it is so widespread. Mapping the 21-cm line provides data to measure galaxy rotation curves (e.g., de Blok et al. 2008), dynamical effects of spiral shocks and bars, and non-gravitational forces such as those driven by stellar winds and supernova remnants (Tamburro et al. 2009). But the hydrogen atoms are much more than test particles. The warm and cool atomic media represent transitional stages in the cycle of galactic chemical evolution. The hot gas driven out of evolved stars and supernova explosions is rich in heavy elements; it eventually cools and is swept up into interstellar clouds from which new stars form. The atomic phase comes after the H II has cooled enough to recombine but before molecule formation has converted most of the H I to H₂,  Fig. 11-1. Although radiation in the 21-cm line is not energetically significant for cooling the gas, it provides an excellent thermometer that shows us how the atomic ISM is distributed over temperature in the full range from 10 to 10⁴ K. The atomic hydrogen is thus the substrate of the Galaxy, at once the background and the sustenance of the Galactic ecosystem.

The 21-cm line which is used to trace the interstellar H I fortuitously lies at a wavelength where the earth's atmosphere is quite transparent and the technology of antennas and receivers is well developed. In particular, both the emission in the line and the absorption by the line of background continuum emission are easy to measure. It is both a complication and a great advantage of this line that both emission and absorption appear in Galactic spectra, often mixed



 Fig. 11-1

The cycle of Galactic evolution. Star formation enriches the interstellar medium with heavy elements through supernova remnants and red giant winds. To form new generations of stars and planets, the hot ejected gas must make its way around the right side of this cycle back to cold, dense conditions where star formation can take place through gravitational collapse. The H I traces two steps, the warm neutral medium (WNM) and the cool neutral medium (CNM) that fill much of the space between the stars

together. Using different kinds of telescopes and observing strategies it is often possible to separate these two effects, and thus to determine both the optical depth and excitation temperature of the H I.

Since its prediction in 1944 and its near-simultaneous detection by three independent research teams in 1951 (reviewed by van de Hulst et al. 1954), the 21-cm line has been one of the most valuable tracers of the Galactic disk. In the 60 years since its detection, this line has inspired theoretical and computational research through several generations. Recent reviews by Kalberla and Kerp (2009) and Furlanetto et al. (2006) give comprehensive discussions of the Galactic and cosmological applications of the H I line, respectively, and the Ferrière (2001) review of the interstellar environment is still an excellent resource. This chapter is not a research review like those; this is an introduction to a few important topics with explanations designed to be accessible to someone who is not an expert in the field. As the subject is the H I disk of the Milky Way, observations of the 21-cm line at low latitudes are necessarily the main source of information, and form the heart of all the topics discussed.

➤ Section 2 is a brief summary of existing instruments and observational techniques and their limitations. In ➤ Sect. 8 some future instruments and the advances they will bring are considered. The radiative transfer equations for the 21-cm line are derived in some detail in ➤ Sect. 3. One of the complications of this line is that it is sometimes optically thick, particularly in low latitude spectra, so the interpretation of emission and absorption spectra is discussed and illustrated. One of the most fundamental observational results at low latitude is the longitude-velocity diagram; this is explained in ➤ Sect. 4, along with the significance of the velocity gradient in setting the brightness of the H I line. ➤ Section 5 considers the H I density as a function of height above the plane, z , and its connection with the random velocity distribution function, in the inner and outer Galactic disk. Starting in ➤ Sect. 6 the theoretical basis for the observed properties of the H I are considered, with thermal equilibrium first. The small-scale structure of the atomic medium is presented in ➤ Sect. 7. But the concept of an interstellar cloud as a structure needs to be considered first.

1.1 Structures in the Atomic ISM

Interpretation of the distribution of optical obscuration and reddening of starlight was pioneered by Chandrasekhar and Munch (1952). Until the mid-1970s the structure in the ISM density field was interpreted as the manifestation of a cloud ensemble, where the variations of the observed properties of the medium (optical depth, column density, or emission measure) represent statistical differences between samples taken from an underlying cloud population as an ensemble. A highly developed treatment of this model as applied to 21-cm emission spectra is given by Mebold et al. (1974). This paper includes a discussion of the correlation expected from one telescope beam area to the next due to the larger clouds in the ensemble that cover several beams. The cloud mass spectrum has a long history, going back to Oort (1955) and Field and Saslaw (1965); a thorough historical review is given by Elmegreen and Scalo (2004). Modern ISM simulations demonstrate that *the very concept of an interstellar cloud can be deceiving*. Local maxima in the gas density come and go in these simulations often without persisting long enough to interact with their environment as coherent structures (Ballesteros-Paredes and Mac Low 2002; Vázquez-Semadeni et al. 1995).

Molecular clouds and cloud complexes certainly do exist; self-gravity is significant at the high densities and low temperatures present in this ISM phase. But in the H I medium self-gravity is generally insignificant, so regions of high density need some external pressure to prevent them dissipating at the sound speed. This is a problem in the far outer MW disk, and in the Magellanic Stream, where the H I is gathered in large structures, but it is hard to trace a confining medium with sufficient pressure to explain their size and apparent age (Strasser et al. 2007).

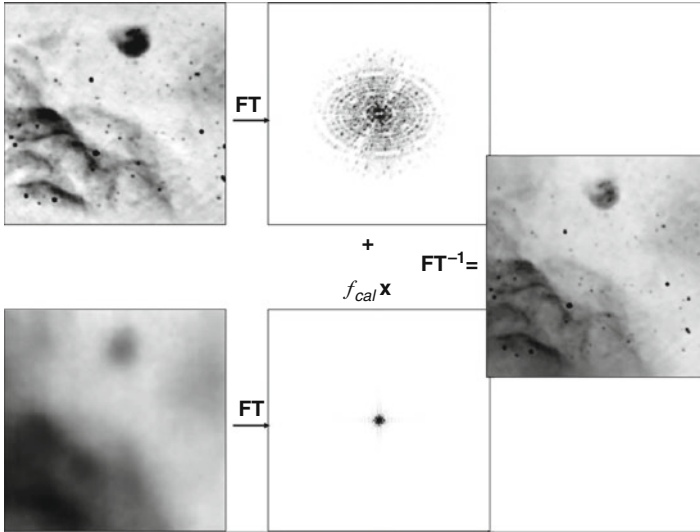
2 Instruments and Techniques for 21-cm Line Observing

Different kinds of radio telescopes are useful to measure emission vs. absorption in the 21-cm line, and within each major category every telescope has its own strengths and weaknesses. As for all telescopes, resolution is an important consideration, but it must be achieved in different ways for emission and absorption observations. Interferometers of various designs are standard for achieving high resolution at radio frequencies, but for Galactic 21-cm line studies they have a fatal weakness in not measuring the zero spacing and very short spacings on the uv plane. (Useful textbooks on radio astronomy instrumentation covering aperture synthesis and uv plane concepts include Taylor et al. 1999, Thompson et al. 1986, and Rohlfs and Wilson 2006.) Since interferometers can only measure baselines longer than the diameters of the individual antennas, they necessarily provide images of the sky brightness that is high-pass filtered, i.e., convolved with a spatial filter that removes Fourier components with low spatial frequencies. These lost low spatial frequencies contain much or even most of the brightness of the line, since at some velocities the entire sky is covered with 21-cm emission. Techniques for combining data from single dish and interferometer telescopes have been perfected that allow the short spacings to be filled in (a good discussion is given by Stanimirović et al. 1999). This is often combined with the mosaicing observing technique to cover large areas of the sky with nearly uniform sensitivity, illustrated in [▶ Fig. 11-2](#).

As always in astronomy, aperture determines sensitivity and hence the integration time needed to reach a given detection limit. But in Galactic 21-cm emission line observations the sensitivity limit that counts is usually the noise in brightness temperature. The noise in flux density applies to unresolved sources, which is important for absorption studies and for detection of very small clouds, but not for the more common goal of making maps of the distribution of the gas in position and velocity.

The brightness sensitivity of aperture synthesis instruments is set by the covering factor of the array antennas; in recent surveys this is about 1 K (rms) for a beam size of $1'$ for a spectrometer channel width of 1 km s^{-1} . Pure single dish surveys by telescopes like Arecibo and Parkes that have covering factor of 1 can go to much higher brightness sensitivity (McClure-Griffiths et al. 2009; Stanimirović et al. 2006). [▶ Figure 11-2](#) illustrates a single field seen with an interferometer mosaic, a single dish, and the combination.

The spatial high-pass filter effect provided by an interferometer or aperture synthesis telescope with baselines longer than about 1 km (fringe separation of $45''$) is helpful to separate the emission from the absorption toward a compact background continuum source. At centimeter wavelengths the sky is rich in extragalactic radio sources. Typically there are about two such sources per square degree brighter than 100 mJy (Condon 1989 and references therein); above 20 mJy the number increases to 10 deg^{-2} . These can be used to get H I absorption spectra, if



■ Fig. 11-2

Combining interferometer and single dish maps. The *upper left-hand* figure is an Australia Telescope Compact Array (ATCA) interferometer mosaic map of Galactic λ 21-cm continuum emission with resolution $\sim 2'$, the *lower left* map is the same region seen with the $14'$ beam of the Parkes single dish telescope. The Fourier transforms of each show the distribution of the emission on the aperture or uv plane. Combining with the proper scale factor and transforming back to the image plane gives the *right-hand panel* (Illustration courtesy of Naomi McClure-Griffiths)

the emission in the direction of the source can be subtracted with good precision. An estimate for the line emission toward a compact continuum source can be made by interpolating spectra taken around the source, but how accurately this interpolation can be done depends on the resolution of the telescope and the angular variations of the emission itself. Small-scale fluctuations in the H I set a limit on the precision of the interpolated emission spectrum, and thus on the resulting absorption spectrum. For a background source with continuum flux density F_c , the optical depth noise, σ_τ , set by emission fluctuations with amplitude ΔT is

$$\sigma_\tau = \frac{\Delta T}{G F_c} \quad (11.1)$$

where G is the antenna gain (K Jy^{-1}) and a typical value for ΔT is 10% of the total emission, T_B , for points separated by a few arc minutes. At low latitudes ($b \sim 0^\circ$) the brightness temperature of the H I line emission can be $T_B \sim 100$ K, so $\Delta T \sim 10$ K. For a 100 m class single dish telescope $G \sim 1 \text{ K Jy}^{-1}$, so these emission fluctuations cause optical depth noise of $\sigma_\tau = 0.1$ for a 100 Jy continuum source. There are only a few continuum sources brighter than 100 Jy in the entire sky at $\lambda 21$ -cm, so single dish telescopes of 100 m diameter or smaller are not able to measure H I absorption in very many directions. Arecibo can do considerably better (Heiles and Troland 2003a and references therein), but to get accurate absorption spectra toward a large number of background sources an array like the Very Large Array (VLA) or even the Very Long Baseline Array (VLBA) is much more effective. A survey with beam size of $1'$ is less confused by emission fluctuations that make the interpolation of the emission toward the continuum

source inaccurate. A background source with flux density of 100 mJy gives a marginally useful absorption spectrum ($\sigma_\tau \sim 0.1$) for a survey with a beam size of $1'$ even at low latitudes where the emission is bright and highly variable. For Galactic latitudes $|b| < 2^\circ$ the combined catalog of absorption spectra from the Canadian Galactic Plane Survey (CGPS, Taylor et al. 2003), VLA Galactic Plane Survey (VGPS, Stil et al. 2006), and Southern Galactic Plane Survey (SGPS, McClure-Griffiths et al. 2005) presented by Dickey et al. (2009) is the most extensive collection currently available. The small-scale structure of both the emission and absorption are important astrophysical questions in their own right, as discussed in more detail in [Sect. 7](#) below.

Since the background sources themselves are mostly a few arc seconds in diameter, the optimum resolution for absorption surveys is about $10''$ – $20''$. With sufficient integration time absorption spectra can be obtained with optical depth rms $\sigma_\tau \ll 0.01$ (Braun and Kanekar 2005; Stanimirović and Heiles 2005) toward a continuum source with flux density ~ 100 mJy with a telescope beam of $\sim 15''$ like that of the VLA, Westerbork Synthesis Radio Telescope (WSRT), or Australia Telescope Compact Array (ATCA). The next generation of cm-wave aperture synthesis telescopes like the Australian Square Kilometer Array Pathfinder (ASKAP, Johnston et al. 2007) will provide a much better catalog of 21-cm absorption spectra as well as spectral line cubes of the brightness temperature of the H I line in emission with unprecedented resolution and sensitivity ([Sect. 8](#) below).

For surveys of 21-cm emission, a parameter more critical than resolution for many applications is the dynamic range of the telescope, in particular its rejection of stray radiation, i.e., emission from other parts of the sky that enter the beam through so-called back sidelobes. Stray radiation is a worse problem for the 21-cm line than for other lines because such a large area of the sky near the Galactic plane is so bright. If there is any blockage of the aperture, e.g., by a subreflector and its supporting legs, then some very weak response over a large part of the sky will be caused by reflections off this blocking structure and from diffraction around it, so that faint traces of any emission that is above the horizon can blend with the spectrum in the direction where the telescope is pointed. This is fatal to studies of H I in the MW halo, or in the outer disk, where the signal is very faint. The unblocked aperture of the Green Bank Telescope alleviates this problem, and the meticulous calibration and correction applied to the Leiden-Argentine-Bonn (LAB) Survey has produced an all-sky atlas that is free of stray radiation to a level of < 0.05 K (Kalberla et al. 2005). More recently the Parkes Galactic All Sky Survey (GASS, McClure-Griffiths et al. 2009, Kalberla et al. 2010) has stray radiation removed to an even deeper level. Coupled with its smaller beam size and high sensitivity, this will be the best survey of the H I sky for many years. The data are available from a web site of the Australia Telescope National Facility.

3 Radiative Transfer in the 21-cm Line

3.1 Brightness Temperature

The energy of a photon, hf , emitted from the hyperfine-split ground state of atomic hydrogen is 5.87×10^{-6} eV, which is small compared with kT_s for any reasonable excitation temperature, T_s (e.g., $\frac{hf}{kT_s} = 0.025$ for $T_s = 2.7$ K, where k is Boltzmann's constant, f is frequency, and h is Planck's Constant). Thus the ratio of the population in the upper hyperfine level (u) to that in the lower level (l) is nearly in proportion to the multiplicity, g , of the levels, $\frac{g_u}{g_l} = 3$. Expanding

the Boltzmann factor to first order gives

$$\frac{n_u}{n_l} = \frac{g_u}{g_l} e^{-\frac{hf}{kT_s}} \simeq 3 \left(1 - \frac{hf}{kT_s}\right) \quad (11.2)$$

This is nearly independent of the excitation temperature, T_s , as long as the atoms are all in the ground state, which will be true if the kinetic temperature, $T_k \ll 10^5$ K for collisional excitation, since the first excited electronic state of H is 10.2 eV above the ground state. In the Milky Way disk T_s is nearly equal to T_k , not only because of the long lifetime of the hyperfine transition against spontaneous deexcitation, but also because of the thermalization of the line via the transitions to and from the first excited state by absorption and emission of Lyman α photons (Furlanetto et al. 2006, [Sect. 2](#)).

The emission coefficient, j_f , gives the energy radiated per unit volume per unit time per unit solid angle per unit bandwidth, as a function of frequency. It is given by the population in the upper level times the photon energy times the spontaneous deexcitation rate, A_{ul} , as

$$j_f \Delta f = \frac{n_u A_{ul} hf}{4\pi} \quad (11.3)$$

where the 4π assumes isotropic emission and Δf is a bandwidth that includes some or all of the photons emitted; Δf is the equivalent width of the emission line if the value of j_f is taken at the line peak and n_u includes all the atoms in the upper level in a unit volume. If Δf is taken to be narrower than the width of the line, e.g., the bandwidth of a single spectrometer channel, then the value of j_f over this bandwidth measures only a subset of n_u restricted to be the density of only those atoms in the upper level whose photons are emitted in the frequency range f to $f + \Delta f$. The distribution of the emission over frequency, which is usually determined by the Doppler shift due to the line of sight component of the velocity of each atom, can be parameterized by a profile function $p(f)$ with units Hz^{-1} , which translates to $p(v)$ with units $(\text{km s}^{-1})^{-1}$. If the line profile is normalized by $\int_{-\infty}^{+\infty} p(f) df = 1$, then using the value of $A_{ul} = 2.884 \cdot 10^{-15} \text{ s}^{-1}$ for the 21-cm transition gives

$$\frac{j_f}{\text{erg s}^{-1} \text{ cm}^{-3} \text{ Hz}^{-1} \text{ ster}^{-1}} = 1.61 \cdot 10^{-33} \left(\frac{n}{\text{cm}^{-3}} \right) p(f) \quad (11.4)$$

with n the total density of H I, $n \simeq \frac{4}{3} n_u$ by [\(11.2\)](#). Good references for the physics of emission and absorption in this and other lines is Spitzer (1977) Sects. 3.3–3.4 and Draine (2011) Chapter 8. Precise values of f and A for the H I hyperfine transition are given by Gould (1994).

If the line is optically thin then the intensity, I_f , measured by a telescope is proportional to the column density of H I atoms, $N \equiv \int_s n ds$ with ds an increment along the line of sight, s , starting at the telescope. At any given frequency, f ,

$$I_f = \int_s j_f ds = 1.61 \cdot 10^{-33} \frac{N}{\text{cm}^{-2}} p(f) \text{ erg cm}^{-2} \text{ Hz}^{-1} \text{ sterad}^{-1} \quad (\text{for } \tau \ll 1) \quad (11.5)$$

thus the column density of hydrogen is the fundamental quantity measured in a 21-cm emission survey. Usually the intensity, I_f , is expressed in terms of the Planck function with an equivalent temperature, called the brightness temperature, T_B , that is the temperature giving I_f as:

$$T_B(f) = \frac{\lambda^2}{2k} I_f = \frac{I_f}{6.20 \cdot 10^{-19} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ Hz}^{-1} \text{ sterad}^{-1}} \quad (\text{for } \lambda = 21.1 \text{ cm}) \quad (11.6)$$

where the Planck function simplifies to the Rayleigh–Jeans approximation because $\frac{hf}{kT} \ll 1$ and the numerical value is for the 21-cm line rest wavelength. Integrating the brightness temperature

over frequency gives the total column density as

$$\frac{\int T_B(f) df}{\text{K Hz}} = 2.60 \cdot 10^{-15} \frac{N}{\text{cm}^{-2}} \quad (\text{for } \tau \ll 1) \quad (11.7)$$

where the integral on the left is taken over the line profile. Using the Doppler shift to convert frequency to the more common units of km s^{-1} gives

$$\frac{N}{\text{cm}^{-2}} = 1.82 \cdot 10^{18} \frac{\int T_B(v) dv}{\text{K km s}^{-1}} \quad (\tau \ll 1) \quad (11.8)$$

because a velocity step of 1 km s^{-1} corresponds to a frequency step of 4.738 kHz at zero redshift. Note that a Maxwellian velocity distribution in one dimension gives a line profile function, $p(v)$, that is a Gaussian with dispersion

$$\frac{\sigma_v}{\text{km s}^{-1}} = 0.0908 \sqrt{\frac{T_D}{\text{K}}} \quad (11.9)$$

where the Doppler Temperature, T_D , would equal the kinetic temperature if there were no random velocities besides those arising from the thermal motions of the individual atoms. Generally T_D is larger than T_s by a factor of 2 or 3, because T_D includes contributions to the line width from bulk random velocities on a range of scales. A better measure of the kinetic temperature is obtained from T_s , if it can be derived from a combination of emission and absorption spectra.

3.2 H I Cloud Masses

The column density can be converted to the hydrogen mass, M_H , by integrating over the solid angle of a cloud, and multiplying by distance, d , squared:

$$M_H = m_H \int_{\text{area}} N_H dA = m_H d^2 \int N_H d\Omega \quad (11.10)$$

with m_H the hydrogen atomic mass, or in solar masses:

$$\frac{M_H}{M_\odot} = 1.46 \cdot 10^4 \left(\frac{d}{\text{kpc}} \right)^2 \int d\Omega \left(\frac{\int dv T_B(v, \alpha, \delta)}{\text{K km s}^{-1}} \right) \quad (11.11)$$

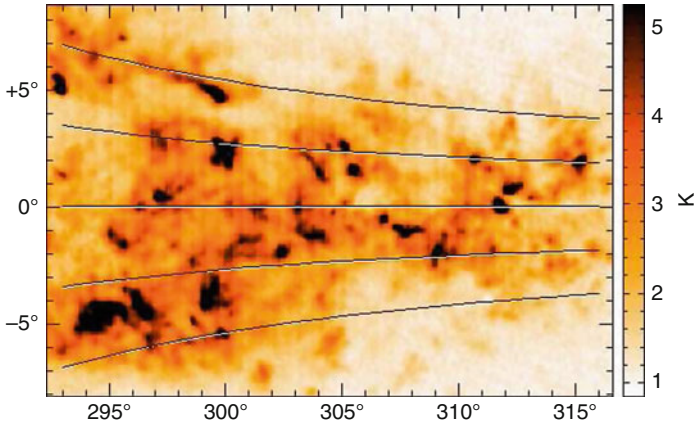
where α and δ can be any coordinates on the sky, and the integral is taken over the solid angle of the emission feature, whatever it is.

If the cloud is unresolved it appears on a map with just the size and shape of the telescope beam; in that case the brightness temperature cannot be measured. The value of T_B on the map is the average over the beam area, while the unresolved cloud must have a higher T_B since its angular size is smaller than the beam. In this case the relevant quantity is the flux density,

$$F = \frac{2k}{\lambda^2} \int_{\Omega_s} T_B d\Omega \quad (11.12)$$

which is measurable even if the source solid angle, Ω_s , is less than the beam solid angle. In units of Janskys ($1 \text{ Jy} = 10^{-23} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ Hz}^{-1}$) this gives

$$\frac{M_H}{M_\odot} = 0.236 \left(\frac{d}{\text{kpc}} \right)^2 \frac{\int F(v) dv}{\text{Jy km s}^{-1}} \quad (11.13)$$



■ Fig. 11-3

Small clouds that are part of the Lockman layer, i.e., gas with scale height of ~ 400 pc. The figure shows T_B , with the scale in K, for a single velocity channel ($\delta v = 0.8 \text{ km s}^{-1}$). The velocity of each spectrum has been shifted by a different amount at each longitude, so that the terminal velocity is at a constant value of -125 km s^{-1} . This map is made at velocity -145 km s^{-1} , i.e., 20 km s^{-1} beyond the cutoff from Galactic rotation. This velocity shifting isolates the population of intermediate velocity clouds ($\sigma_v \approx 20 \text{ km s}^{-1}$) along the circle of tangent points (● Sect. 4.1). The curved lines trace $z = 0, \pm 200, \pm 400$ pc at the tangent point distance for each longitude

where the integral is over the flux density spectrum, giving total line flux in units of Jy km s^{-1} . Some example small clouds seen in the GASS survey are shown in ● Fig. 11-3. Most are resolved with the Parkes telescope (beam size $14'$ FWHM). Typical angular diameters are $20'$, which translates to 23 pc at a distance of 4 kpc. These are similar to the clouds cataloged by Ford et al. (2008). For example, a typical brightness temperature of $T_B = 5 \text{ K}$ over a velocity width of 5 km s^{-1} gives column density $4.6 \cdot 10^{19} \text{ cm}^{-2}$. Integrated over a circle with diameter 20 pc this translates to H I mass of $115 M_\odot$.

3.3 Optical Depth

The 21-cm line is not always optically thin by any means, so the assumption implicit in ● Sects. 3.1 and ● 3.2 that $\tau \ll 1$ often breaks down. The optical depth as a function of velocity, $\tau(v)$, is relatively easy to measure on some lines of sight, and this can be used to determine the excitation or “spin” temperature of the transition, T_s , as a function of frequency or radial velocity. The absorption coefficient, κ_f , multiplied by I_f gives the amount of energy removed from the radiation field by the gas in a unit volume, per unit time, bandwidth, and solid angle, so κ_f is defined by

$$I_f \kappa_f \Delta f = \frac{hf}{4\pi} I_f (n_l B_{lu} - n_u B_{ul}) \quad (11.14)$$

where the Einstein coefficients for absorption (B_{lu}) and stimulated emission (B_{ul}) are given by

$$g_l B_{lu} = g_u B_{ul} = g_u A_{ul} \frac{c^2}{2hf^3} = 2.04 \cdot 10^5 \frac{\text{cm}^2}{\text{erg s}} \quad (11.15)$$

(see Spitzer 1977, [▶ Sect. 3.2](#)). Using the Taylor expansion of ([● 11.2](#)) gives

$$\kappa_f \Delta f = \frac{hf}{4\pi} n_l B_{lu} \left(1 - \frac{n_u g_l}{n_l g_u} \right) \simeq \frac{hf}{4\pi} n_l B_{lu} \left(\frac{hf}{kT_s} \right) \quad (11.16)$$

here the bandwidth Δf is the equivalent width in absorption if the peak value of κ_f is used with the total number density of atoms in the lower state, n_l . Alternatively, if a subset of the atoms are selected based on a range of frequencies, Δf , or the corresponding range of radial velocities, Δv , in the line profile, then $\kappa_f \Delta f$ gives the absorption resulting from those atoms alone. For a line profile function $p(f)$ as in ([● 11.4](#)), the result for the absorption coefficient is

$$\frac{\kappa_f}{\text{cm}^{-1}} = 2.60 \cdot 10^{-15} \left(\frac{n}{\text{cm}^{-3}} \right) \left(\frac{T_s}{\text{K}} \right)^{-1} p(f) \quad (11.17)$$

and integrating the absorption coefficient along a line of sight (s) gives the optical depth:

$$\tau_f = \int_s \kappa_f ds = 2.60 \cdot 10^{-15} \int_s \left(\frac{T_s}{\text{K}} \right)^{-1} \frac{n ds}{\text{cm}^{-2}} p(f) \quad (11.18)$$

where this integral is similar to that giving the column density in ([● 11.5](#)), except that here the density is weighted by the inverse of the excitation temperature. The integral of optical depth over frequency, $\int \tau_f df$, is proportional to this line-of-sight integral:

$$\int_s \frac{n}{T_s} ds = 1.82 \cdot 10^{18} \text{ cm}^{-2} \text{ K}^{-1} \left(\int \frac{\tau(v) dv}{\text{km s}^{-1}} \right) \quad (11.19)$$

where on the right-hand side the integral of the absorption line profile over frequency has been converted to units of velocity based on the Doppler shift. The term in parentheses on the right-hand side of ([● 11.19](#)) is the equivalent width of the absorption line in velocity units.

Lines of sight through the Milky Way sample gas moving at a range of velocities, thus the line profile in either emission or absorption can be a complicated superposition of features corresponding to gas at different distances. [▶ Figure 11-4](#) shows a pair of 21-cm line spectra in emission and absorption taken at $(l, b) = (36.06, +0.36)$, with the frequency axis translated to radial velocity as usual. In this direction Galactic rotation projects on the line of sight to spread the gas in the inner Galaxy over a velocity range of 0–100 km s⁻¹, and gas in the outer Galaxy on the far side of the solar circle appears at negative velocities. The strikingly different appearance of the emission and absorption is typical of H I spectra taken in any direction, at high or low latitudes, and it reflects the different temperature weighting in the line of sight integrals in ([● 11.5](#)) and ([● 11.18](#)). In this example, some of the absorption lines show as peaks or shoulders on the emission spectrum, while others show as dips due to absorption by cold clouds in front of most of the emission.

The fundamental equation of radiative transfer is

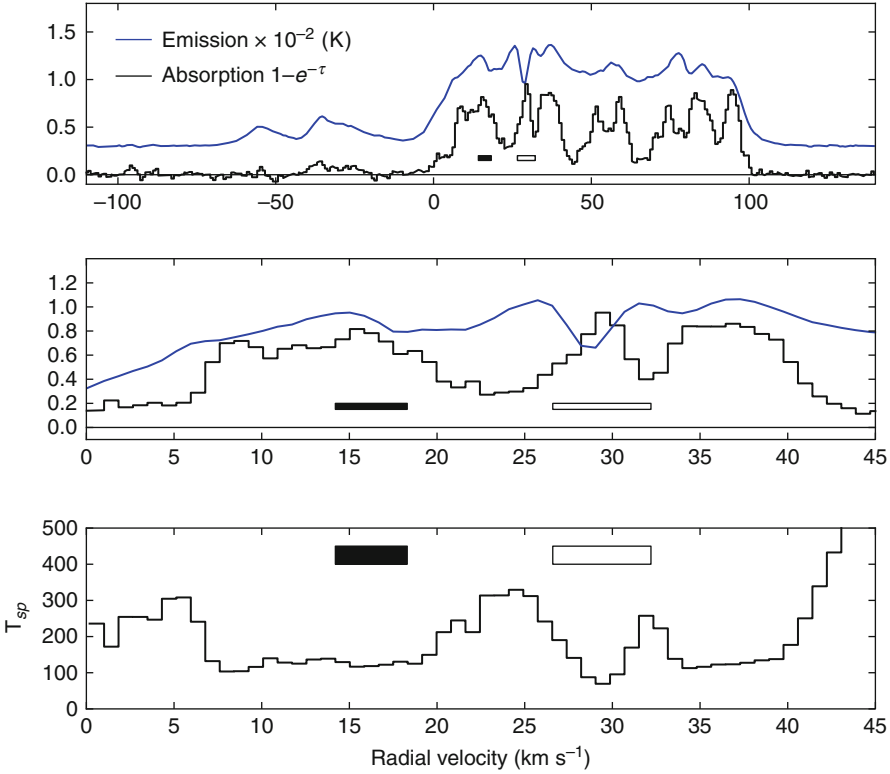
$$\frac{dI_f}{dx} = j_f - \kappa_f I_f \quad (11.20)$$

where dx is a step following the radiation along the line of sight toward the telescope, $dx = -ds$, so $d\tau_f = -\kappa_f dx$ and

$$\frac{dI_f}{d\tau_f} = \frac{j_f}{\kappa_f} + I_f \quad (11.21)$$

Bringing the I_f to the left side, multiplying by $e^{-\tau}$ and integrating gives

$$I(0) = e^{-\tau(d)} I(d) + \int_0^{\tau(s)} \frac{j_f}{\kappa_f} e^{-\tau} d\tau \quad (11.22)$$



■ Fig. 11-4

Emission and absorption spectra toward $(l, b) = (36.06, +0.36)$ from the VGPS (Dickey et al. 2009). The broader, *smooth curve* shows the emission, scaled as $T_B \cdot 10^{-2}$ K, while the fine, *stepped line* shows the absorption, $1 - e^{-\tau}$, both plotted as functions of radial velocity, v , measured relative to the Local Standard of Rest (LSR). In the *top frame* only, the emission is offset by +0.2 (20 K) for clarity. In the second and third panels the velocity scale is expanded to show details of some of the line features between 0 and +45 km s $^{-1}$. At $v \approx 27 - 32$ km s $^{-1}$ there is a clear self-absorption (HISA) feature, where T_B dips from 100 K down to 60 K, corresponding to a narrow absorption line with $\tau \approx 3$, i.e., $1 - e^{-\tau} \approx 0.95$. This velocity range is marked with an *open box*. Most of the other absorption features correspond roughly in velocity with emission peaks, e.g., at $v_{\text{LSR}} = 15$ km s $^{-1}$ where $T_B = 95$ K and $\tau = 1.7$, marked with a *solid box*. The *lower panel* shows the emission divided by the absorption: $T_1 \equiv \frac{T_B}{1 - e^{-\tau}}$ (☛ 11.24)

where the integral is taken from the telescope ($s = 0$) along the line of sight through the medium, and $I(d)$ is the background emission originating at some large distance $s = d$ behind the absorbing cloud. For an isothermal, homogeneous cloud this gives simply

$$T_B = T_s (1 - e^{-\tau}) + T_{\text{bkg}} e^{-\tau} \quad (11.23)$$

where I is converted to brightness temperature using (☛ 11.6) and the background intensity, $I(d)$, becomes T_{bkg} . All the variables in (☛ 11.23) are functions of frequency f or Doppler velocity, v , across the spectrum.

Observing the emission and absorption allows derivation of T_s as

$$T_1(\nu) = \frac{T_B(\nu)}{1 - e^{-\tau(\nu)}} \quad (11.24)$$

$$\simeq \frac{\int_s n ds}{\int_s n T_s^{-1} ds} \quad (\tau \ll 1) \quad (11.25)$$

where the second line applies in the limit of low optical depth, so $(1 - e^{-\tau}) \simeq \tau$, and the subscript 1 is used to denote that this observed quantity corresponds to the physical excitation temperature, T_s , if there is gas at only one temperature contributing to the line of sight integrals at a given velocity, ν . In spectra taken at high latitudes where there is little blending of multiple line components, $T_1(\nu)$ generally shows a local minimum at the velocity where the optical depth is highest, and rises on either side of the line center. This is the natural effect of the H I existing at a range of temperatures in and around an absorbing cloud. Since the warmer gas has a wider line profile, $p(\nu)$, the line wings are more dominated by warm gas than the center velocity. Differences in the profile shape between the emission and absorption can best be described by a superposition of gas at different temperatures, contributing differently to the brightness temperature and optical depth.

If emission and absorption by gas at two temperatures, T_a and T_b , overlap in velocity, and if neither region is very optically thick, then the temperatures add as a harmonic mean weighted by the column densities, N_a and N_b , in (🔗 11.25) to give

$$T_1 = \frac{N_a + N_b}{\left(\frac{N_a}{T_a}\right) + \left(\frac{N_b}{T_b}\right)} \quad (\tau_a, \tau_b \ll 1) \quad (11.26)$$

where these are all functions of velocity across the spectrum. If either or both of the optical depths are not small, then the juxtaposition of the two regions is important in determining the resulting brightness temperature. If the cooler gas is closer than the warmer gas, then the combined brightness will be lower than if the warmer gas is closer, as

$$T_B = T_a (1 - e^{-\tau_a}) + T_b (1 - e^{-\tau_b}) e^{-\tau_a} \quad (11.27)$$

where the radiation from the more distant cloud, b , is partially absorbed by the closer cloud, a .

Combining emission and absorption spectra by assuming a single temperature in each velocity channel, as in (🔗 11.24), is clearly an over-simplification, since T_1 can vary by an order of magnitude or more from the line center to the line wings of a single cloud. Since there are only two measurables at each velocity, T_B and τ , there is not enough information to determine multiple pairs of numbers, N_a and T_a , N_b and T_b , for each of two or more regions of gas contributing to the emission and absorption spectra. In a more realistic picture, the temperature is a continuously varying function of position in an interstellar cloud, and not necessarily a monotonically or smoothly varying function. So a better way of interpreting the combination of emission and absorption is needed, that goes beyond the assumption of a single excitation temperature as in (🔗 11.24). The next simplest thing to do to simplify (🔗 11.26) is to assume two temperatures are present in the gas, but the warmer one, T_w , is so high that the absorption from the warm gas, τ_w , is negligible compared to that of the cool gas, τ_c . The brightness temperature contribution of the warm gas,

$$T_{B,w} = T_w (1 - e^{-\tau_w}) \simeq T_w \tau_w \quad (11.28)$$

is not small, but it contributes only to the numerator in (11.26), thus

$$T_1 = \frac{N_c + N_w}{\frac{N_c}{T_c}} \quad (11.29)$$

with T_c the excitation temperature of the cooler gas. This gives the fraction of gas in the cool state relative to the total

$$f_c \equiv \frac{N_c}{N_c + N_w} = \frac{T_c}{T_1} \quad (11.30)$$

which can be determined if a value for T_c is assumed. Numbers from about 50 to 70 K are typical for T_c , but it can be as low as 20 K or as high as several hundred K (Dickey et al. 2003; Heiles and Troland 2003b). Using a value of $T_c = 60$ K generally gives a result for f_c in the range 0.25–0.45 for the H I in the Milky Way disk, and similar numbers for nearby spiral galaxies, with the Magellanic Clouds having a lower value of $f_c \sim 0.2$ –0.25 (Mebold et al. 1997).

To make a more accurate determination of the excitation temperatures along a line of sight requires fitting profile shapes to the emission and/or absorption across the velocity range of the spectral features (Dickey et al. 2003). An alternative method is to fit Gaussian features to the absorption spectrum, and then fit Gaussians to the emission spectrum assuming that some of the line centers and widths in emission correspond to the absorption features (Heiles and Troland 2003b). At low latitudes, there is typically a jumble of blended spectral features in emission and absorption covering the range of velocities allowed by Galactic rotation for the inner Galaxy. Sometimes a distinct absorption line component has an emission component centered at nearly the same velocity with similar linewidth. On Fig. 11-4 two examples are marked by solid and open boxes. These can each be interpreted as an individual cloud with excitation temperature T_c and optical depth τ_c , plus foreground and background gas contributing to the emission and absorption as:

$$T_B = T_{s,f} (1 - e^{-\tau_f}) + T_c (1 - e^{-\tau_c}) e^{-\tau_f} + T_{s,b} (1 - e^{-\tau_b}) e^{-(\tau_f + \tau_c)} \quad (11.31)$$

where f designates the gas in the foreground and b designates the gas in the background. If the optical depth of the cloud, τ_c , is narrow in velocity so that the contributions of the foreground and background gas are nearly constant over the velocity range of the cloud, then the cool phase temperature, T_c , of the cloud can be estimated by rewriting (11.31) as

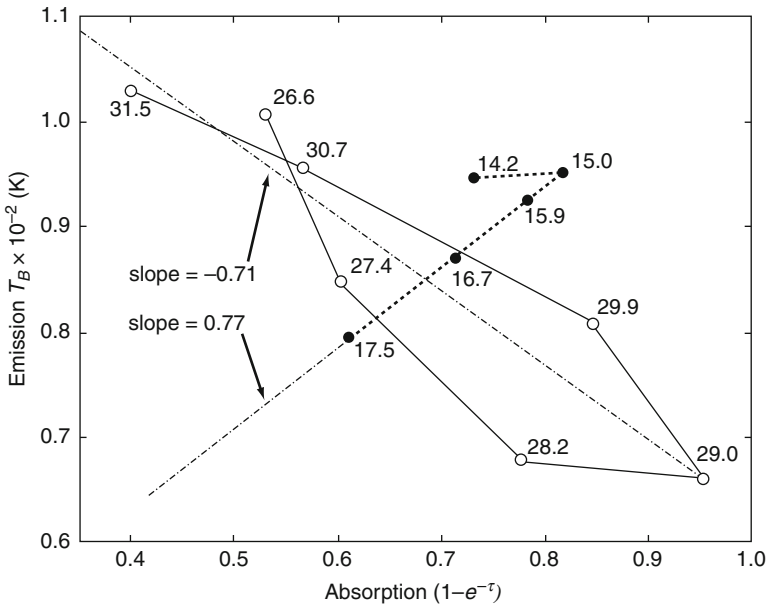
$$T_B = (T_{B,f} + T_{B,b} e^{-\tau_f}) + [e^{-\tau_f} (T_c - T_{B,b})] (1 - e^{\tau_c}) \quad (11.32)$$

where the brightness temperature of the foreground has been rewritten as $T_{B,f} = T_{s,f} (1 - e^{\tau_f})$ and similarly for the background brightness temperature, $T_{B,b}$.

Under the assumption that the foreground and background contributions to the emission and absorption spectra are roughly constant in velocity over the linewidth of a separate absorption feature, (11.32) gives a nearly straight line segment on a plot of T_B vs. $(1 - e^{-\tau})$, as in Fig. 11-5. This figure has points for the emission and absorption in each velocity channel for the two velocity ranges marked by filled and open boxes on Fig. 11-4. Both boxes include discrete absorption features; for one the emission also shows a peak at the center velocity of the absorption line, for the other the emission shows a dramatic dip at the velocity of the absorption. On Fig. 11-5 the channels in both ranges show nearly linear relationships between emission and absorption; the slopes correspond to the term in square brackets, $[e^{-\tau_f} (T_c - T_{B,b})]$, in (11.32), i.e., 77 and -71 K respectively. To solve for T_c requires knowing how much of the emission and absorption come from the foreground and background of the cloud. For example,

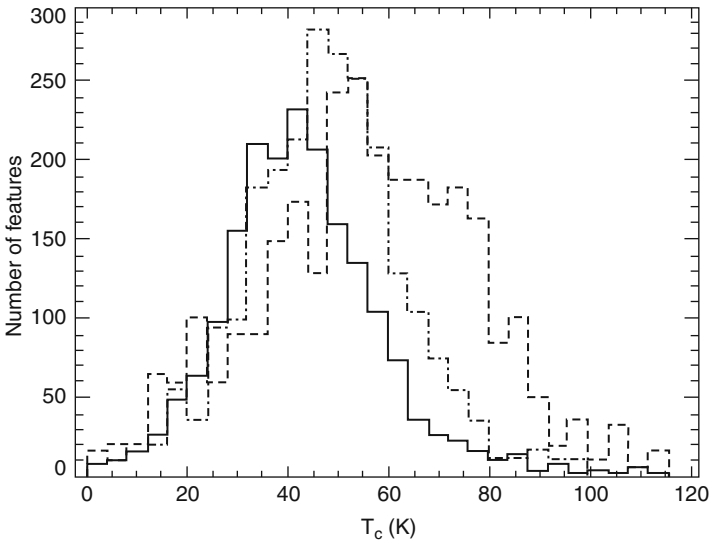
the cloud with slope of 77 K shows off-line values of $[(1 - e^{-\tau}), T_B] = [0.6, 80 \text{ K}]$ measured at $\nu = 17.5 \text{ km s}^{-1}$. In the extreme case where the cloud is in front of all the other gas, $e^{-\tau_f} = 1$ and $T_{B,b} = 80 \text{ K}$, so the slope of 77 K indicates $T_c \simeq 157 \text{ K}$. In the other extreme, with all the other gas on the line of sight in front of the cloud, $T_c \simeq 192 \text{ K}$, since $e^{-\tau_f} = 0.4$ and $T_{B,b} = 0$. A more reasonable assumption is that the foreground and background have equal amounts of emission and absorption, respectively, so that $\tau_b = \tau_f = 0.46$ and $T_{B,f} = T_{B,b} = 49 \text{ K}$. Then $T_c \simeq 171 \text{ K}$. This is close to the harmonic mean between 157 and 192; in this case the cloud has peak optical depth $\tau = 0.69$, measured at $\nu = 15.0 \text{ km s}^{-1}$ where the total $(1 - e^{-\tau}) = 0.8$. Each spectral channel has velocity width of 0.8 km s^{-1} in this case, so the column density is $N = 1.7 \cdot 10^{20} \text{ cm}^{-2}$ in the central channel where $T_c = 171 \text{ K}$ and $\tau_c = 0.69$. For the cloud as a whole N is about $6 \cdot 10^{20} \text{ cm}^{-2}$.

The second example on [Fig. 11-5](#) corresponds to the open box on [Fig. 11-4](#), this has slope -71 K and off-line values of $[(1 - e^{-\tau}), T_B] = [0.5, 100 \text{ K}]$ measured at $\nu = 26$ and $\nu = 31 \text{ km s}^{-1}$ on either side of the absorption component centered at 29.0 km s^{-1} . If the cloud is in front of all the rest of the gas at this velocity on the line of sight, then $T_c = 29 \text{ K}$. If the cloud is in front of just half of the gas, with equal contributions to τ in the foreground and background, i.e., $\tau_f = \tau_b = 0.35$, and equal brightness temperature in emission, $T_{B,f} = T_{B,b} = 59 \text{ K}$, then there is no valid solution ($T_c = -42 \text{ K}$), and there is certainly no solution for the case where the cloud



■ Fig. 11-5

Fitting for T_c in the two spectral features marked with boxes in [Fig. 11-4](#). Consecutive spectral channels from the velocity ranges marked by the filled and open boxes are here marked by *filled* and *open circles*, with the values of ν_{LSR} indicated in km s^{-1} . In both cases the emission and absorption follow a nearly linear relationship. The slope of a straight line segment that fits the data gives a better indication of the cool gas temperature, T_c , using [\(11.32\)](#), than the simple T_1 [\(11.24\)](#) drawn on [Fig. 11-4](#)



■ Fig. 11-6

The distribution of cool gas temperatures, T_c , using (◆ 11.32) to separate blending in velocity. The three curves represent samples taken from the CGPS (*solid*), SGPS (*dash*), and VGPS (*dot-dash*) from Strasser (2006, p. 74). The number of features for the SGPS and VGPS have been scaled up by a factor of 5. There is a long tail of higher temperature clouds ($T_c > 100$ K) that are not included because they are hard to distinguish in emission spectra, either as emission peaks or self-absorption dips. The feature corresponding to the solid box on ◆ Fig. 11-4 is typical of this class of blended cloud with a relatively warm temperature indicated


is behind all the other gas ($T_c \rightarrow -\infty$). Thus in this case 29 K is a strict upper limit on T_c , and the cloud is probably cooler, since some of the emission must be in the foreground.

There are many variations on the technique of spectral analysis described in the preceding paragraphs. The simplest is to fit Gaussians to the emission and absorption, if there are unblended features clearly corresponding between the two spectra. It is often important to include the effects of diffuse continuum emission, particularly when the slope of T_B vs. τ is negative, which corresponds to H I self-absorption (next section). Taking large samples of spectra and many distinct absorption lines in each, and assuming 50% for the typical fraction of gas in the background, gives distributions of cool cloud temperatures, T_c , as shown in ◆ Fig. 11-6.




3.4 H I Self-absorption



It is common at low Galactic latitudes that cold foreground clouds absorb the emission from gas behind, as in the open box line component on ◆ Figs. 11-4 and ◆ 11-5. This effect is often called H I self-absorption (HISA), although it is not self-absorption in the normal radiative-transfer sense, because the absorbing cloud may be far away from the background spectral line emission. HISA can be distinguished from velocity structure in the emission spectrum due to fluctuations in the H I density distribution by its narrow spectral features

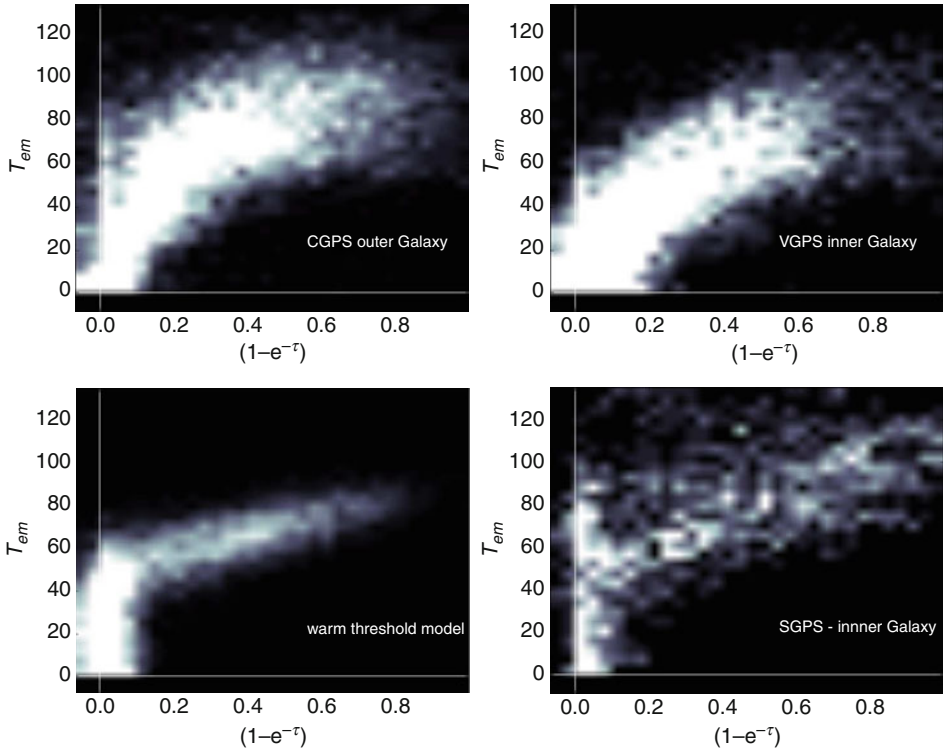
(Gibson et al. 2005a; Kavars et al. 2005). An extensive survey of H I associated with cold molecular clouds (Krco et al. 2008) has shown that the HISA correlates well with molecular line tracers of the cold gas. In the L134 and L1757 molecular clouds the cool phase H I excitation temperature, T_c , is consistent with the temperature measured with the ^{12}CO line, with values between 14 and 19 K. This atomic gas is a trace constituent in the molecular cloud, probably caused by cosmic ray dissociation of H_2 . Although the CO emission in these nearby molecular clouds generally corresponds in velocity with dips in the H I spectra, which Krco et al. call H I narrow line self-absorption or HINSA, it is not the case that all HISA features correspond with molecular lines, whether of CO or OH or other molecules. Most HISA features do not arise in the dense molecular clouds ($n_{\text{H}_2} > 10^3 \text{ cm}^{-3}$) that dominate the CO emission at low latitudes. Thus although most nearby molecular clouds show HISA, most HISA does not arise in molecular clouds. Matching HISA features against CO surveys can be an effective way to determine whether molecular clouds in the inner Galaxy should be placed at the near or far kinematic distance (Anderson and Bania 2009; Kolpak et al. 2003; Roman-Duval et al. 2009).

On much larger scales there are more diffuse HISA features that can be traced over several degrees in the CGPS (Gibson et al. 2005b). Estimating the column density of the gas responsible suggests that some 25–30% of the H I mass of the Milky Way is in the form of cool clouds. At low latitudes, the typical brightness temperature of the aggregate H I emission is 25–75 K; thus in emission surveys the cool clouds appear sometimes as positive features and sometimes as HISA. And sometimes they are effectively invisible against a background whose brightness temperature is nearly equal to the excitation temperature of the gas in the cloud, which would give a horizontal line segment with slope zero on a plot like  [Fig. 11-5](#).

3.5 The Relationship Between H I Emission and Absorption

A more coarse use of the plot of T_B vs. $(1 - e^{-\tau})$ is to grid the distribution of points taken for each velocity channel in many pairs of spectra, as shown in  [Fig. 11-7](#) for samples taken from three different low-latitude surveys. The points do not follow a constant excitation temperature, so the mixture of warm and cool gas, f_w and f_c , must change with optical depth, $(1 - e^{-\tau})$. All three samples show similar behavior, which can be fitted by a model where some warm gas, N_w , is associated with the cool clouds that appear as features in absorption. The simplest such model gives the fit labeled “warm threshold model” on the lower left panel of  [Fig. 11-7](#). In this case the optical depth is close to zero (with random noise added) until the emission brightness temperature reaches $T_{\text{threshold}} = 50 \pm 10 \text{ K}$ corresponding to $N_w = 9.1 \cdot 10^{19} \text{ cm}^{-2}$ in a spectral channel of width 1 km s^{-1} ( [Fig. 11-8](#)). Above this column density in the model, the rest of the gas is in the cool phase with $T_c = 70 \text{ K}$, giving the line with constant slope. (The cool gas absorbs the radiation from half of the warm gas that is assumed to lie behind it, so the slope is not 70 K but about 45 K instead.)

The structure of H I clouds is certainly more complicated than this simple model suggests. A thoughtful consideration of the effects of geometry is given by Liszt (1983). In both atomic and molecular clouds, there is good evidence for a hierarchy of density and temperature variations over a wide range of scales ( [Sect. 7](#) below). But the agreement of the observational data from the three surveys shown on  [Fig. 11-7](#) is strong evidence that shielding by WNM is needed in most environments for the existence of CNM. A column density of $\sim 10^{20} \text{ cm}^{-2}$ would give extinction of only about 0.05 mag in the visual, but in the ultraviolet and soft X-rays which



■ Fig. 11-7

Plotting the emission vs. absorption for many channels of many spectra in different low-latitude emission-absorption surveys. The grey scale shows the density of points in spectra from the inner Galaxy (only) in the first and fourth quadrants, from the VGPS (*upper right*) and SGPS (*lower right*). The *upper left* panel shows the density of points in the outer Galaxy (only) from the CGPS. This is similar to [Fig. 5](#) of Strasser and Taylor (2004). The *lower left* panel shows the result of the Monte Carlo model described in the text

are a major source of heating through photoelectric electrons from grains, the effect of this column density can be significant in reducing the heating rate (Wolfire et al. 2003, see [Sect. 6](#) below).

4 The Longitude–Velocity Diagram and the Velocity Gradient

The rotation curve of the Milky Way is the most fundamental function for study of Galactic structure and dynamics. It is measured using many different tracers; this section considers what the H I shows, starting with some well-known equations relating radial velocity to distance in a disk in circular rotation. There is good evidence that there are departures from circular rotation, including coherent patterns on large scales, but the first topic to consider is the simpler situation of a disk in pure cylindrical rotation with some rotation curve, $\Omega(R)$, i.e., the angular velocity as a function of galactocentric radius, R .

Observing any tracer of radial velocity, V_r , and assuming circular rotation only, the rotation curve can be measured at each longitude in the inner Galaxy by finding the extreme or terminal velocity of the line emission. In this situation two semicircles in the disk are responsible for our estimate of the rotation curve. These semicircles are the places where our line of sight is tangent to a circle of constant R . All these points together, the so-called locus of sub-central points, make a circle centered on the point half way between the sun and the Galactic center. Plots of the H I rotation curve show irregularities on small and intermediate scales (arc minutes to degrees) because the curve is not an azimuthal average over the entire inner Galaxy, but a measure of the rotation velocity, Θ , only along these two semicircles from the sun to the Galactic center.

4.1 Kinematics in a Circularly Rotating Disk

For a disk in circular, cylindrical rotation the observed radial velocity, V_r , of a point at longitude l and latitude b is the difference between the projection of the velocity of that point along the line of sight and the projection of the velocity of the Local Standard of Rest (LSR) along the line of sight. This gives the familiar

$$V_r = R_o [\Omega(R) - \Omega_o] \sin(l) \cos(b) \quad (11.33)$$

which is derived using the law of sines in the triangle made by the line of sight from the sun to the point of interest, and the lines joining each of these points to the Galactic center. An excellent review including a derivation of (11.33) is given by Burton (1988). Here R_o and Ω_o are the solar circle radius and the LSR angular velocity of rotation, respectively. It is more common to work with the rotation curve in velocity units (km s^{-1}), thus $\Theta_o = R_o \Omega_o$ is the LSR velocity around the Galactic center. R_o and Θ_o are the fundamental scale parameters of Galactic structure. Equations in the next paragraphs of this section use upper case variables (e.g., S , V_r) for lengths and velocities in absolute units (kpc, km s^{-1}), and lower case variables (s , v) for lengths and velocities in units of R_o and Θ_o .

As long as the angular velocity of rotation, $\Omega(R)$, does not rise with Galactic radius, R , meaning that the rotation velocity $\Theta(R)$ does not rise faster than for a solid body, i.e., $\Theta(R) \propto R$, then the radial velocity of points along a line of sight through the inner Galaxy reaches an extremum at the point closest to the Galactic center (the sub-central or tangent point). This extremum, and all other velocities for points inside the solar circle, are positive in the first quadrant ($0^\circ \leq l \leq 90^\circ$) and negative in the fourth quadrant ($270^\circ \leq l \leq 360^\circ$), while velocities of gas outside the solar circle are negative in the first and second quadrants, positive in the third and fourth. Note that the assumption that $\Omega(R)$ is a monotonically decreasing function is reasonable, since as long as the average mass density inside R decreases with radius, R , then the angular velocity will also decrease with R . A solid body rotation curve corresponds to a constant density distribution with spherical symmetry.

In a recent analysis of tangent point velocities, McClure-Griffiths and Dickey (2007) suggest that the rotation curve $\Theta(R)$ can be approximated by the simple function

$$\Theta(R) = \Theta_o \left(K_1 + K_2 \frac{R}{R_o} \right) \quad (11.34)$$

with $K_1 = 0.887$ and $K_2 = 0.186$. Formal errors in K_1 and K_2 are about 1 and 4%, respectively, but note that Levine et al. (2008) find values of 0.844 and 0.212 for these parameters, in the context of a model that also fits values of the change in Θ with height above the plane, Z . Note that

$(K_1 + K_2) \neq 1$, which is inconsistent at the solar circle. The radial extent over which (11.34) is valid is given by $r \equiv \frac{R}{R_o}$ in the range $0.35 < r < 0.95$. Over this range, departures of the measured tangent point velocities from the predictions of this simple polynomial fit have amplitudes 5–10 km s⁻¹.

For a line of sight through the disk ($b = 0^\circ$) inside the solar circle at longitude l , the observed radial velocity of a point at Galactic radius $R = rR_o$ is

$$V_r = \left(\frac{K_1}{r} + K_2 - 1 \right) \Theta_o \sin(l) \quad (11.35)$$

from substitution of (11.34) into (11.33). Scaling by Θ_o gives

$$v \equiv \frac{V_r}{\Theta_o} = \left(\frac{K_1}{r} + K_2 - 1 \right) \sin(l) \quad (11.36)$$

Using the law of cosines in the triangle formed by the sun, the Galactic center, and a point on the line of sight at distance $S \equiv s \cdot R_o$ from the sun, gives

$$s = \cos(l) \pm \sqrt{r^2 - \sin^2(l)} \quad (11.37)$$

or

$$r = \sqrt{1 + s^2 - 2s \cos|l|}. \quad (11.38)$$

Solving (11.36) for r and combining with (11.38) gives a simple kinematic distance formula

$$s = \cos(l) \pm \sqrt{\cos^2|l| - \left[1 - \frac{K_1^2}{\left(\frac{v}{\sin(l)} + 1 - K_2 \right)^2} \right]} \quad (11.39)$$

where v is the measured radial velocity in units of Θ_o and s is the line of sight distance from the sun in units of R_o . Unfortunately kinematic distances in the inner Galaxy are bi-valued, meaning that a given radial velocity corresponds to two points along the line of sight, spaced on either side of the tangent point by a distance given by the square root term in (11.39).

4.2 Rotation Curve Models

Besides the simple linear fit of (11.34), there are many other functions that have been proposed as descriptions of the Milky Way rotation curve, a list is given by McClure-Griffiths and Dickey (2007). These mostly give kinematic distance relations more complicated than (11.39); but they are valid over a larger area of the disk. A simple function that gives a fit to the terminal velocity data about as good as the linear fit is a Brandt curve that has functional form:

$$\frac{\Theta(R)}{\Theta_m} = \frac{R}{R_m} \left[\frac{1}{3} + \frac{2}{3} \left(\frac{R}{R_m} \right)^n \right]^{-\frac{3}{2n}} \quad (11.40)$$

with parameters $\Theta_m = 228 \text{ km s}^{-1}$, $R_m = 8.5 \text{ kpc}$ and $n = 0.7$. These parameters are quite close to those derived by Brandt (1960), although he was working with a different scale factor ($R_0 = 10 \text{ kpc}$ which was the IAU standard until the early 1980s). Most importantly, the decrease in Θ outside the solar circle expected from a Brandt curve does not match the data, either in the Milky Way or other spiral galaxies. A fair approximation to the outer Galaxy rotation curve is simply a constant $\Theta(R) = \Theta_0$ for $R > R_0$ (Kalberla and Kerp 2009), although a slight rise

in the rotation velocity with R as claimed by Brand and Blitz (1993) is possible. A sharp rise in $\Theta(R)$ in the bulge region, R between 0.5 and 3 kpc ($0.06 < r < 0.35$) cannot be due to azimuthally symmetric circular motion, as discussed by Burton and Liszt (1993), because in the range 0.4–1.0 kpc it drops faster than the prediction given the stellar mass density of the bulge. Burton and Liszt suggest that the true circular velocity would be a function slowly rising with radius, continuous with the more easily measured rotation curve for $r > 0.35$. More recently, stellar surveys show that dynamics in the bulge are strongly influenced by a stellar bar (Howard et al. 2009; Weiner and Sellwood 1999), but more work needs to be done on H I and other ISM tracers to reconcile the bar mass model with the dynamics of the gas.

4.3 Modeling the Longitude–Velocity Diagram

Surveys of H I at low latitudes in the Milky Way are unlike surveys of most other tracers because the atomic gas is so widespread. Thus rather than considering the H I emission to be coming from discrete clouds that can be distinguished and measured individually, like molecular clouds, it is better to start from a paradigm where the gas simply fills space uniformly. In fact the volume filling factor of the H I is considerably less than one, probably varying from 25% to 60%, but it is instructive to consider what the emission would look like in a spectral line from a hypothetical ISM tracer that is optically thin with constant density and hence constant emissivity in a circularly rotating disk.

For an optically thin, uniformly distributed spectral line tracer, the brightness would be proportional to the inverse of the velocity gradient along the line of sight, $\frac{dV_r}{dS}$, with V_r from (11.33). This is because spectrometer channels or any series of equal steps in frequency or radial velocity, (ΔV_r), will map into unequal length steps ΔS along the line of sight, with the step length given by

$$\Delta S = \left| \frac{dS}{dV_r} \right| \Delta V_r \quad (11.41)$$

Converting to dimensionless quantities s and v , the derivative is

$$\frac{ds}{dv} = \frac{\Theta_o}{R_o} \frac{dS}{dV_r} = \left(\frac{\frac{ds}{dr}}{\frac{dv}{dr}} \right) \quad (11.42)$$

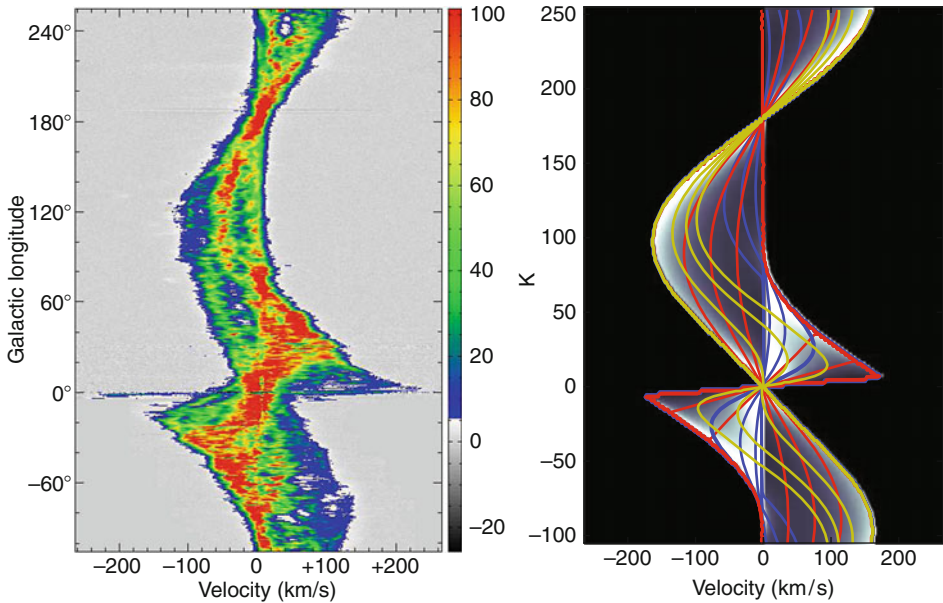
Using (11.34), this becomes

$$\left| \frac{ds}{dv} \right| = \frac{r^3}{K_1 \sin |l| \sqrt{r^2 + \cos^2(l)} - 1} \quad (11.43)$$

where this derivative will be positive on the near side of the tangent point in the first quadrant and on the far side of the tangent point in the fourth quadrant, and negative otherwise. To evaluate the derivative, substitute for r using

$$r = \frac{K_1}{\frac{v}{\sin(l)} + 1 - K_2} \quad (11.44)$$

from (11.36) as was already used to get (11.39). The longitude-velocity diagram for this hypothetical emission line will resemble the right-hand panel of Fig. 11-8, which plots the velocity gradient from (11.43) using the values of K_1 and K_2 given above.



■ Fig. 11-8

Longitude–velocity diagrams for the Milky Way H I line. On the *left* is a complete picture, made mostly with the Green Bank Telescope by the author in 2001. Data from the Parkes telescope observations for the SGPS is used to fill the gap $-90^\circ < l < -20^\circ$. On the *right* above is a model based on the prediction of (☛ 11.43) and (☛ 11.44). *Contour lines* indicate points with $R = 0.1, 0.3, 0.6, 1.2, 1.5, 2,$ and 4 times R_0 in *red*, and in *blue* and *yellow* the same values for distance from the sun, S . The velocity gradient prediction matches the overall behavior of the H I brightness fairly well in some areas, particularly in the inner Galaxy and around the anti-center ($l = 180^\circ$). The greatest mismatch is in the far outer Galaxy, where the H I density is dropping rapidly with radius, so the constant density model greatly overestimates the brightness

The high-velocity boundary of the emission in the inner galaxy in both panels on ☛ Fig. 11-8 corresponds to the projected rotation curve at the tangent points minus $\Theta_0 \sin l$, which gives the terminal velocity at each longitude. The departures from a smooth curve in the real data indicate the amplitude of deviations from circular rotation along the locus of tangent points. Looking in more detail at gas beyond the terminal velocity in the inner Galaxy, Kang and Koo (2007) show that there are occasional, broad wings in the emission that are sometimes coincident with known supernova remnants.

The structure in the brightness at velocities between zero and the terminal velocity is caused by several physical effects. The simplest is inhomogeneities in the density distribution of the gas. More important are irregularities in the velocity field, similar to those that cause small-scale structure in the terminal velocity. Small departures from circular rotation can cause large deviations from the smooth brightness distribution predicted by the velocity gradient. Finally, self-absorption by cool gas clouds in the foreground imposes modulation or variations in the brightness of the background emission. This effect is particularly significant in the inner

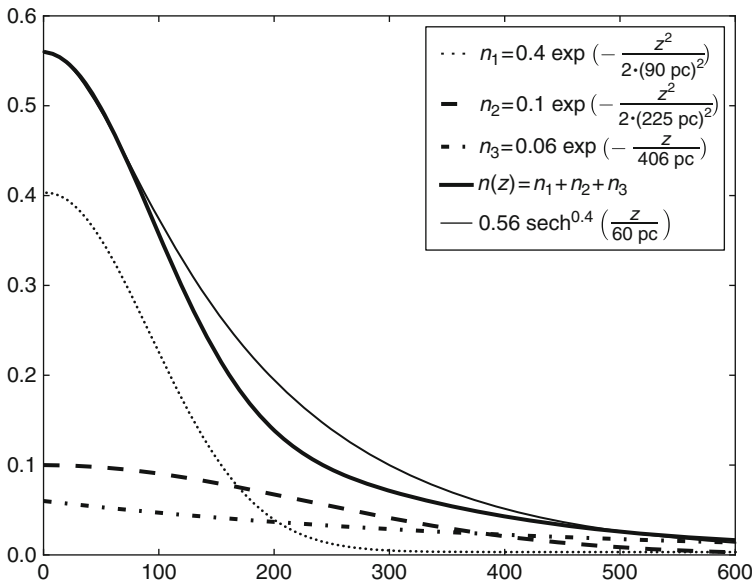
Galaxy where the velocity–distance relation is double-valued, so that clouds on the near side are superposed on background emission at the same velocity.

5 The Structure of the H I Disk

5.1 The z Distribution in the Solar Neighborhood

By studying the brightness of the 21-cm line as a function of latitude near the tangent point circle and the solar circle, the distribution of the density of the H I, $n(z)$, as a function of height above the plane, z , can be determined. In the solar neighborhood, $n(z)$ is best described by the sum of three functions, a Gaussian with dispersion $\sigma_1 \simeq 90$ pc and midplane density $n_1 = 0.4$ cm^{-3} , another Gaussian with dispersion $\sigma_2 \simeq 225$ pc and midplane density $n_2 = 0.1$ cm^{-3} , plus an exponential with scale height $h \simeq 400$ pc and midplane density $n_3 = 0.06$ cm^{-3} (Dickey and Lockman 1990, shown in [Fig. 11-9](#)). This aggregate $n(z)$ distribution has a half-width-to-half-maximum (HWHM) of 126 pc.

The random velocity distribution of the H I in the plane measured beyond the terminal velocity in the inner galaxy is also made up of three separate components, with velocity widths $v_{rms} = \sqrt{\langle v^2 \rangle}$ of 4.5, 9, and 18 km s^{-1} (McClure-Griffiths and Dickey 2007). The CNM-rich,



■ Fig. 11-9

The z distribution of H I density, $n(z)$, from Dickey and Lockman (1990), based on data from Lockman et al. (1986), see also Malhotra (1995). The two Gaussian plus one exponential components are indicated separately, with their sum shown by the wider line. A $\text{sech}(z)$ function of the form given in [\(11.45\)](#) that fits the low z and high z portions of $n(z)$ is shown. In spite of the contribution of the intermediate width Gaussian term, n_2 , the $\text{sech}(z)$ function still overestimates the density at midrange z (150–300 pc) by about 25%

cool clouds correspond to the narrow z component and the narrow v component, while the two broader components are mostly all WNM. However, there must be a significant fraction of WNM associated with the CNM clouds in the narrow component, since the abundance and depth of absorption lines relative to the emission suggests that the H I is roughly 70% WNM, 30% CNM overall.

The exponential and Gaussian distributions for the three components of $n(z)$ can be associated through the more general function

$$n(z) = n_0 \left[\operatorname{sech} \left(\frac{z}{z_0} \right) \right]^\alpha \quad (11.45)$$

(Celnick et al. 1979; Kalberla et al. 2007). This function approaches a Gaussian with dispersion $\sigma_z = \frac{z_0}{\sqrt{\alpha}}$ for $z \ll z_0$ and an exponential with scale height $h = \frac{z_0}{\alpha}$ for $z \gg z_0$. For an infinite, self-gravitating gaseous disk in hydrostatic equilibrium the $\operatorname{sech}(z)$ function solves the force balance equation:

$$K_z = \frac{1}{n m_H} \frac{dP}{dz} \quad (11.46)$$

where K_z is the acceleration due to gravity as a function of z , i.e., $K_z = \frac{\text{Force}}{\text{mass}}$ analogous to $g = 980 \text{ cm s}^{-2}$ on Earth, and P is the total pressure, including bulk motions that contribute to the momentum flux in addition to the microscopic thermal motions, and including non-gas components such as the magnetic field and the cosmic rays. These can be parameterized

$$\frac{P}{m_H} = \eta n \langle v^2 \rangle \quad (11.47)$$

where $\eta = \sqrt{1 + \alpha + \beta}$ with α the ratio of the magnetic pressure to the gas pressure, and β the ratio of the cosmic ray pressure to the gas pressure (Parker 1966).

The H I density does not have to satisfy (11.45) because throughout most of the disk the contribution of the atomic gas to the gravitational potential is small compared with that of the stars. Taking the K_z function as a given, a back-of-the-envelope calculation shows the rough equilibrium between the H I scale heights and random velocity distributions as measured beyond the terminal velocity. An extremely simplified approximation for the gravitational mass distribution of the disk at the solar circle is a single layer with constant density, ρ_0 , and half-width, z_h . Then for $|z| \ll z_h$ the K_z function is simply:

$$K_z = -4\pi G \rho_0 z \quad (11.48)$$

with G Newton's constant. It is convenient to convert G to units that fit the scale of the problem: $G = 4.3 \cdot 10^{-3} \left(\frac{\text{km}}{\text{s}} \right)^2 \left(\frac{\text{pc}}{M_\odot} \right)$. For $|z| \gg z_h$ the K_z curve due to the disk flattens to a constant,

$$K_z = -2\pi \Sigma G \quad (11.49)$$

with Σ the mass surface density, $\Sigma = 2\rho_0 z_h$.

If the z dependence of the pressure comes only from variation of the density in (11.49), i.e., $\langle v^2 \rangle$, α , and β are all independent of z , then a Gaussian H I density function, $n(z)$, gives for the derivative $\frac{dP}{dz}$ in (11.46)

$$\frac{1}{n m_H} \frac{dP}{dz} = - \frac{z}{\sigma_z^2} \langle v^2 \rangle \eta \quad (11.50)$$

with σ_z the dispersion of the Gaussian $n(z)$ function. So for $|z| \ll z_h$ as in (11.49), the width of the gas layer and the mean-square velocity are related by

$$\eta \langle v^2 \rangle = 4\pi G \rho_0 \sigma_z^2 \quad (11.51)$$

For example, if $\sqrt{\langle v^2 \rangle} = 7 \text{ km s}^{-1}$, $\sigma_z = 90 \text{ pc}$, and $\rho_0 = 0.11 \text{ M}_\odot \text{ pc}^{-3}$ then $\eta = 1$.

In the case of a gas layer with high velocity dispersion and large scale height, an exponential distribution is appropriate,

$$n(z) = n_0 e^{-\frac{z}{h}} \quad (11.52)$$

with h the scale height, and the pressure derivative becomes

$$\frac{1}{n m_H} \frac{dP}{dz} = -\eta \frac{\langle v^2 \rangle}{h} \quad (11.53)$$

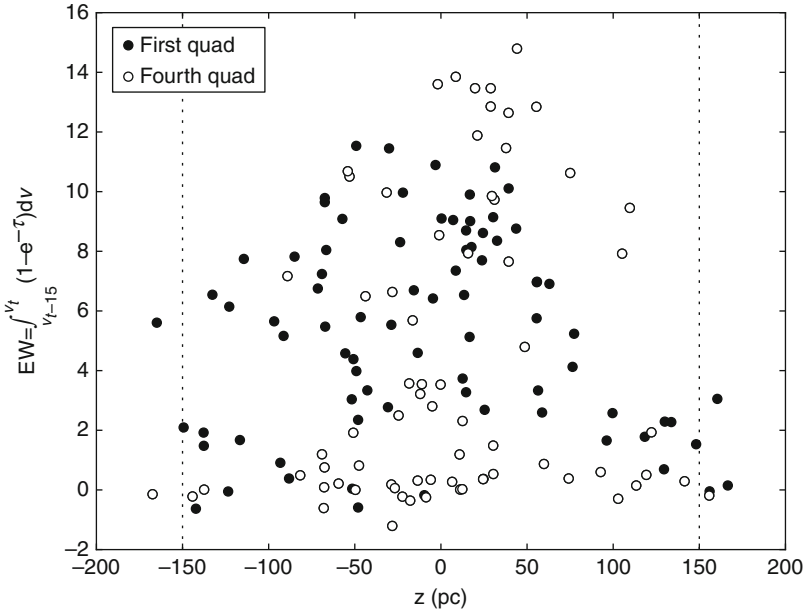
using the same simplifying assumptions about the pressure that were used in the preceding paragraph to get (11.50). Using this with the high z limit for K_z (11.49) gives

$$\eta \langle v^2 \rangle = 2\pi \Sigma G h \quad (11.54)$$

which gives $\eta = 1$ for $\sqrt{\langle v^2 \rangle} = 25 \text{ km s}^{-1}$, $h = 400 \text{ pc}$, and $\Sigma = 58 \text{ M}_\odot \text{ pc}^{-2}$.

While the simple calculation above illustrates the relationship between z velocity dispersion and scale height for the dynamical components of the H I, it does not do justice to the many studies of K_z measured over different areas of the disk using different kinds of stars (e.g., Holmberg and Flynn 2004; Kuijken and Gilmore 1989). Using H I as a dynamical tracer has the advantage that the gas is smoothly distributed throughout the inner and outer disk, but it has the disadvantage that in general accurate distance measurement is not possible. So the function $n(z)$ is determined primarily along the tangent point circle, rather than being an average over a large area of the disk, and the values of $\langle v^2 \rangle$ for the different components are determined at the tangent points for random velocities in the plane of the disk (Lockman and Gehman 1991). The assumption that the velocity ellipsoid is a sphere, i.e., that the random velocity component in the z direction is the same as those in the radial and circumferential directions that can be measured at the tangent points, is reasonable given that collisions in the gas can isotropize the distribution of cloud velocities. But there is little direct observational evidence for this isotropy, although the fact that face-on galaxies generally show velocity dispersion of 7 km s^{-1} or more is strong indirect evidence (e.g., Rownd et al. 1994). The larger question of how the random velocities are generated, and how they relax into an equilibrium distribution, is still controversial (see Sect. 7 below).


The CNM and WNM are differently distributed in z , with the narrow ($\sigma_z = 90 \text{ pc}$) layer rich in CNM while the thicker layers are almost all WNM. This can be seen in Fig. 11-10, that shows height above or below the plane, z , at the tangent point for each line of sight toward a strong background source in the SGPS and VGPS, with the integral of the corresponding absorption spectrum, $(1 - e^{-\tau})$, over velocity just for the last 15 km s^{-1} before the terminal velocity. This velocity range selects gas near the tangent point (typically within $\pm 1 \text{ kpc}$ line of sight distance from the tangent point). The overall distribution shows that higher optical depth is found closer to the midplane, and for $|z| > 100 \text{ pc}$ the optical depth integral is typically half its average value at $z = 0$. This result needs more analysis, since the z limits of the surveys are lower than the scale height of the Lockman Layer, i.e., the exponential component of the $n(z)$ distribution, as indicated by the dashed lines on Fig. 11-10. Other large absorption surveys have found similar results, e.g., Belfort and Crovisier (1984) derive $\langle |z| \rangle = 100 \text{ pc}$ for the CNM.



■ Fig. 11-10


The distribution of the equivalent width measured over the last 15 km s⁻¹ of the allowed velocity range before the terminal velocity in the inner Galaxy, i.e., $EW = \int_{v_t-15}^{v_t} (1 - e^{-\tau}) d|v|$. The x-axis is the height above or below the plane, z , of the line of sight at the tangent point circle. The *dashed lines* show the approximate survey limits of the VGPS and SGPS. The increase in the width of the distribution in z with decreasing EW indicates that the width of the H I disk increases as the temperature of the clouds decreases. The coldest CNM gas ($EW > 7.5$ km s⁻¹) has a dispersion $\sigma_z = 47$ pc, while the more optically thin and hence warmer gas ($EW < 7.5$ km s⁻¹) has $\sigma_z = 82$ pc

5.2 The Outer Galaxy

The surface density, Σ_H of the atomic gas implied by the $n(z)$ function of the last section is set by the average H I density at $z = 0$, assumed to be $n(0) = 0.57$ cm⁻³, and the scale heights of the different functions. For the two Gaussian plus one exponential model of  Fig. 11-9 the H I surface density is 4.2 M_⊙ pc⁻². This matches recent determinations of Σ_H , as a function of radius in the outer Galaxy (Kalberla and Dedes 2008; Levine et al. 2006). Inside the solar circle the surface density of the H I is roughly constant at this value, although the molecular hydrogen surface density increases with decreasing radius as an exponential with scale length about 3 kpc (Ferrière 2001, Fig. 1). Estimates of the molecular scale length can be biased by the rapid drop in the molecular gas surface density on the outer edge of the molecular ring ($4 < R < 5.5$ kpc), and the overall normalization depends on uncertain quantities like the conversion factor between CO and H₂ column densities. In spite of these uncertainties, it is evident that the surface density of H₂ is greater than that of H I in the molecular ring, probably by a factor of at least 2, but that the molecular surface density decreases with radius fast enough outside the ring that it drops below that of the atomic gas somewhere inside the solar circle. Outside the solar circle, where the ISM surface density is dominated by the H I, the exponential decrease continues, with a

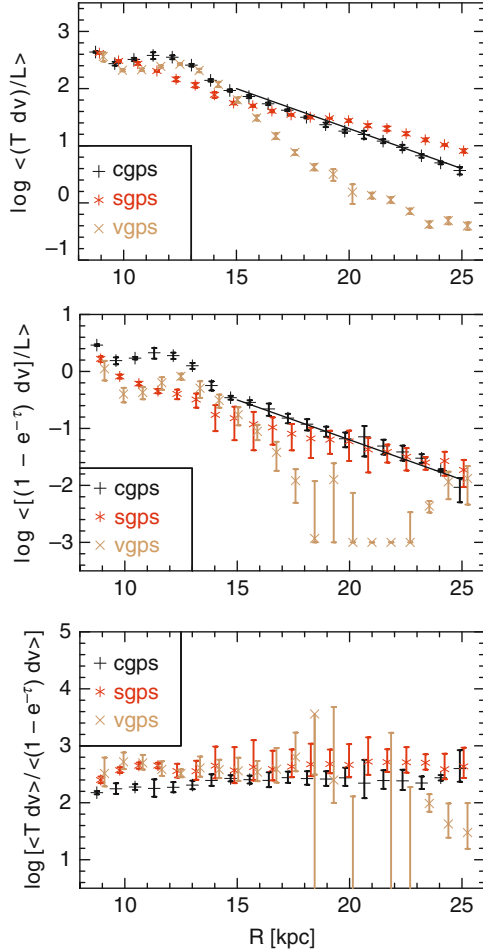
scale length of about 3.75 kpc from about 12 to 30 kpc (Kalberla and Dedes 2008, Fig. 5). Inside the molecular ring, $R < 3.5$ kpc, in the region dominated by the stellar bulge, the H I surface density drops by at least a factor of 2. It rises again inside $R = 0.5$ kpc (Ferrière et al. 2007).

Outside the solar circle, and particularly for $R > 12$ kpc, the H I disk is released from confinement by the thin, flat gravitational potential set by the stars because the stellar surface density of the thin disk has dropped by a factor of 3 or more from its value at the solar circle. The outer disk shows a strong warp, i.e., a departure of the z -centroid of the gas from a flat plane, and a strong flare, i.e., a rapid increase in the scale height with R . Beyond $R = 15$ kpc the gas disk is confined primarily by its own self-gravity and by the dark matter halo. The disk thickness widens to a HWHM of 2 kpc or more at $R = 30$ kpc (Kalberla et al. 2007; Levine et al. 2006). The warp begins as an antisymmetric pattern, but beyond $R = 20$ kpc it becomes very one-sided. For example, at $R = 16$ kpc, in the longitude range $\sim 40^\circ$ to $\sim 80^\circ$, which translates to galactocentric azimuth, ϕ , in the range $60^\circ < \phi < 110^\circ$, the z centroid of the H I rises to positive z . It reaches a maximum of about $z = 1.5$ kpc at azimuth ϕ between 80° and 90° ($l = 58^\circ$). On the other side of the Galactic center at $R = 16$ kpc, there is a nearly symmetric negative warp with peak $z = -1$ kpc at $260^\circ < \phi < 270^\circ$ ($l = 294^\circ$). (Note that ϕ is defined to increase in the same direction as longitude, with zero in the same direction as $l = 0^\circ$, i.e., directly away from the sun starting at the Galactic center.) But at radii $R > 20$ kpc the H I in the fourth quadrant returns to the plane of the inner disk, so that at $R = 28$ kpc the warp shows only a positive peak, with amplitude about 5.5 kpc, still at azimuth $80^\circ < \phi < 90^\circ$, but there is no corresponding negative peak on the other side. The warp can be decomposed into complex Fourier components of the function $z(\phi)$; these components are functions of R (Binney and Merrifield 1998). The first three components give an adequate description the shape of the warp (Binney and Merrifield 1998; Kalberla et al. 2007). Inside of $R = 20$ kpc the warp is nearly sinusoidal in ϕ , so that the first Fourier term dominates.

Looking at the absorption as a tracer of the cool gas in the outer Galaxy, there is an exponential decrease of opacity with radius similar to that of the mean H I density (Dickey et al. 2009; Strasser et al. 2007). Just as it is inside the solar circle, the absorption in the outer Galaxy is confined to a thin disk, with scale height less than that of the emission.  Figure 11-11 shows the emission and absorption per unit line of sight distance, s , which are proportional to the average density $\langle n \rangle$, and to the average of the density divided by the spin temperature, $\langle \frac{n}{T_s} \rangle$. Dividing the average emission by the average absorption, both per unit path length, gives the average excitation temperature, $\langle T_1 \rangle$. It is surprising how constant the spin temperature is with R . All three surveys show $\langle T_1 \rangle \simeq 300$ K to at least $R = 17$ kpc, and in the SGPS and CGPS, where the warp is not a problem, the spin temperature continues with this value to $R = 22$ kpc.

6 Thermal Equilibrium in the H I

Clear evidence that the interstellar H I has a wide range of temperatures came from the first interferometer absorption survey by Clark (1965). The H I excitation temperature and energy balance in the ISM had already been considered theoretically by Field (1959), he soon extended this work to an analysis of the requirements for equilibrium between heating and cooling (Field et al. 1969, hence FGH). At that time the dominant cooling lines had not yet been detected from the ISM, but they were understood theoretically. Even at relatively low temperatures the fine structure lines of CII, CI, and OI in the far infrared are easily excited by collisions, and



■ Fig. 11-11

The radial dependence of the emission and absorption in the three Galactic plane surveys, from Dickey et al. (2009). The warp at longitudes between 30° and 90° in the VGPS and part of the CGPS causes the plane to shift out of the survey area for $R > 16$ kpc, which is why both the emission and absorption drop below the survey limits in the outer part of the VGPS plots. The ratio of the two is T_1 , which is nearly constant with R with a mean value of about 300 K. The *line* indicates a radial exponential with scale length 3.1 kpc

their radiation carries energy out of the medium. These are among the brightest lines at any wavelength emitted by spiral galaxies, but most of their luminosity comes from environments of high mass star formation, called photon dominated or photo-dissociation regions (PDRs, Hollenbach and Tielens 1997). In the diffuse atomic medium the same lines carry away most of the energy, but the total flux is much lower.

The processes that heat the diffuse medium were less obvious in the 1960s and 1970s than the ones that cool. Photoelectric heating by grain absorption of ultraviolet photons that eject energetic electrons was proposed as a significant heating process by Draine (1978). This was recognized to be the dominant heating process only in the late 1980s (Wolfire et al. 1995), when it was appreciated that there is a population of very small grains, called polycyclic aromatic hydrocarbons (PAHs). But the basic model of FGH was not much changed by the changing heating process, since the shape of the equilibrium function comes from the fact that most cooling processes depend on collisions in the gas, and these have a rate proportional to density squared, while most heating processes depend on density to the first power. So the temperature dependence of the cooling, the “cooling curve” (Dalgarno and McCray 1972, Fig. 2) sets the shape of the equilibrium curve in the space of density and temperature. This curve has a dramatic rise for kinetic temperature above about 10^4 K because collisional excitation of the Lyman transitions of hydrogen begins to make a significant contribution. There is a similar rapid rise with temperature below about 100 K, when collisions can begin to excite CI, OI, and the other fine structure lines of the most abundant heavy elements.

Changing axes in the plot of heating–cooling balance to density and pressure, FGH showed how gas at two temperatures (WNM and CNM) could coexist in thermal and pressure equilibrium (• Fig. 11-12). A modern treatment such as Wolfire et al. (2003) considers hundreds of different processes that could contribute to heating and/or cooling depending on the pressure, the radiation field, and the metallicity in the local environment. Yet the conclusion is very similar to that of FGH: There is a range of ISM pressures in which a WNM and a CNM can both be in thermal equilibrium, and that range is $1,500 < \frac{P}{k} < 10^4 \text{ cm}^{-3} \text{ K}$, with the upper limit typically closer to $\sim 3,000 \text{ cm}^{-3} \text{ K}$ except in special conditions. The neutral medium can exist in thermal equilibrium at higher and lower pressures, but at lower pressures it should be all in the WNM phase, and at higher pressures it should be all in the CNM phase.

As the evolution of supernova remnants in the ISM became better understood in the 1970s, it was clear that there must be a wide range of interstellar pressures; at any given point the pressure is set by the age of the most recent supernova remnant (SNR) to pass. This was built into a theory of the dynamical ISM by McKee and Ostriker (1977). In this theory the WNM and CNM are associated in clouds that are ablated when they are engulfed by young SNRs, but the survivors later grow by condensation of the hot, ionized medium inside old SNRs. In this paradigm the H I has a small filling factor, but still a significant fraction of the mass of the ISM, and its properties depend strongly on the supernova rate. The theory was revised considerably to take account of the venting of hot gas into the halo (Norman and Ikeuchi 1989) and the clustering of supernovae (Slavin and Cox 1992). These revisions result in an increased abundance of WNM over CNM, and a larger filling factor for the H I overall, compared with the McKee–Ostriker model. Since the mid 1990s, computer simulations of the ISM have become so powerful and sophisticated that the dynamical relationship between WNM, CNM, and the other ISM phases can be modeled in detail (de Avillez and Breitschwerdt 2007; Vázquez-Semadeni et al. 1995).

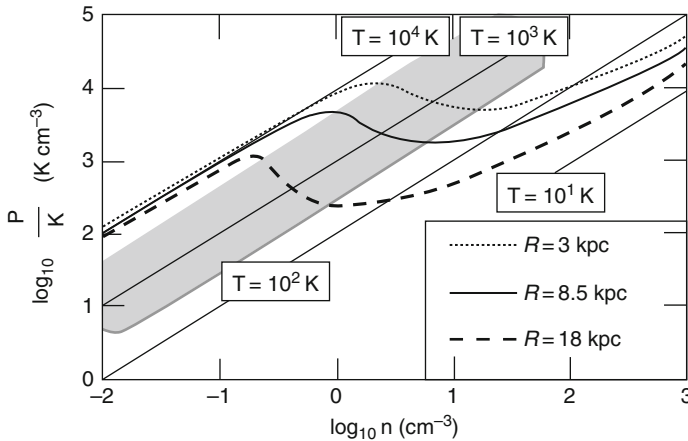
Observations of 21-cm emission and absorption in various environments help guide the theory. The existence of CNM in what should be low-pressure environments such as in the outer disk, the halo, and even in the Magellanic Stream, is particularly challenging. In a plane-parallel disk model such as that discussed above in • Sect. 5.1, the pressure is set by the K_z curve and the gas density. If the gas were isothermal, then

$$P(z) = kTn(z) = kTn(0)e^{\frac{m\Phi(z)}{kT}} \quad (11.55)$$

where $\Phi(z)$ is the gravitational potential, which is -1 times the integral of K_z (► Sect. 5.1). In the Milky Way disk, as in the Earth's atmosphere, the pressure vs. z function is complicated by variations in the temperature with z , but with the added complication that the supernovae are constantly injecting supersonic blast waves that briefly raise the pressure, sometimes by two orders of magnitude from its average value. The SNRs also cause occasional pressure drops below the average, when the old remnants cool rapidly while still at low density. The importance of the SNRs for pressurizing the medium is dramatically illustrated by the large shells and chimneys that are abundant in H I maps of the Milky Way and many other galaxies, such as that in ► Fig. 11-13.

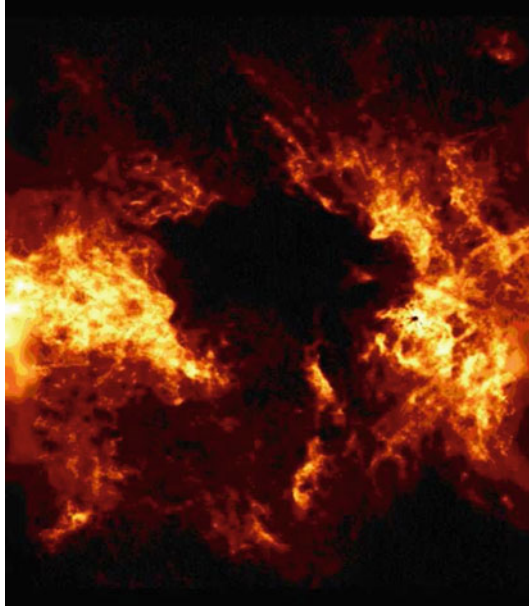
The coexistence of the WNM and CNM is robust, both phases exist in a wide range of environments, even where there are no young stars to cause SNRs. Theoretically this requires that the pressure be in the range on ► Fig. 11-12 where the two phases can coexist. More generally, the pressure must pass all the way through this range occasionally in order to force the gas from one phase to the other. The timescale for heating or cooling to drive the gas from one equilibrium phase to the other becomes a critical quantity. When the gas is far from the equilibrium curve, or when it is at an unstable point on the curve, then it will move toward a stable equilibrium at the same pressure. The heating or cooling time is

$$t_c = \frac{\frac{3}{2}nkT}{\Gamma - n^2\Lambda} \quad (11.56)$$



■ Fig. 11-12

The curve of equilibrium between heating and cooling, derived by FGH, with modifications described in Wolfire et al. (2003). The shape of the equilibrium curve changes with Galactic radius because both the heating and cooling rates depend on the local environment, primarily through the uv radiation field intensity and the abundance of small grains, respectively. If the slope of the curve is negative, the gas is unstable to runaway heating or cooling, the unstable range is indicated by the *shading*. The minimum pressure of the cool phase and the maximum pressure of the warm phase are set by the local maximum and minimum of each curve on the left and right edges of the unstable range. Lines of constant temperature are indicated



■ Fig. 11-13

A large Galactic chimney, GSH277+00+36 (McClure-Griffiths et al. 2003). The combined effect of multiple, overlapping SNRs that generate high pressure and temperature has driven the H I out of a region of the disk many hundred parsecs across. The hot gas breaks out of the surrounding cool gas when the shell opens up into the halo above and below the disk


where the numerator is the heat content of the gas, and in the denominator Γ is the heating rate and $n^2\Lambda$ the cooling rate, both with units $\text{erg cm}^{-3} \text{s}^{-1}$. (Often the denominator is given as just $n^2\Lambda$ for t_c or just Γ for t_w , because once the gas is away from equilibrium one or the other process typically dominates by a large factor.) Values for t_c range from a few times 10^4 years for the CNM to a few times 10^6 years for the WNM. This time is critical in determining whether or not the gas can return to equilibrium after the passage of one SNR before the next one comes. The possible detection of a considerable amount of gas in the forbidden temperature range of a few hundred to a few thousand K (Heiles and Troland 2005) suggests that the H I medium is pushed out of equilibrium so frequently that a significant amount is always passing through the unstable regime. Recent ultra-sensitive studies with the VLA (Dwarakanath et al. 2002), WSRT (Braun and Kanekar 2005), and Arecibo (Stanimirović and Heiles 2005) have not found much intermediate temperature gas on lines of sight to very strong background sources. Instead, these very sensitive absorption spectra suggest that tiny cores of CNM are common wherever there is WNM, even at high latitudes where the line of sight has less than 10^{20} cm^{-2} total column density. These tiny column density CNM clouds ($N_{\text{CNM}} \sim 10^{18} \text{ cm}^{-2}$) may be transient phenomena, as predicted by some simulations. They may be evaporating remnants of larger CNM structures recently disrupted by a young SNR. Or they may be long-lived, possibly as the initial condensations that will eventually grow into a large CNM cloud.

A deeper question is whether the ISM pressure is set by neither the supernova rate nor the disk gravity but by the magnetic field. With observed values of the B field strength in the CNM


of $\sim 6 \mu\text{G}$ (Heiles and Troland 2005), the magnetic pressure would exceed the gas pressure by a factor of three. If the magnetic field dominates the gas pressure then the dynamics of the ISM become much more interesting. Velocity fields in the gas couple to magneto-acoustic wave modes in the B field, and instabilities can drive the field and the cosmic rays that follow it out of the disk into the halo (Parker 1966). If the pressure is effectively independent of temperature then the thermal stability calculation changes. Finally, if the magnetic field dominates the dynamics then the relationship between the small-scale structure in the density and velocity fields to the structure at larger scales can be analyzed as a spectrum of magneto-acoustic waves, rather than a simple Kolmogoroff cascade of turbulent energy from large to small scales where it is dissipated by viscosity (Ferrière et al. 1988).

7 Small-Scale Structure in the H I Medium


7.1 The Spatial Power Spectrum



Structures like the interstellar chimney in  Fig. 11-13 show how variations in the density and velocity field of the H I are generated. Often maps of the ISM show fluctuations on sizes of a few parsecs or less, that do not make a coherent pattern that can be interpreted as the result of a past event like a SNR. Such small-scale density variations can be analyzed using statistical tools to characterize the density and velocity variations as a random process. Measuring the statistics of the H I density leads to a correlation function,

$$R_n(\vec{r}, s) = \langle n(\vec{r}_1) n(\vec{r}_1 + \vec{r}) \rangle \quad (11.57)$$

where s is the distance along the line of sight, n is the density as a function of position, \vec{r}_1 the offset \vec{r} is taken in the plane of the sky, and the averaging $\langle \rangle$ taken over all \vec{r}_1 is effectively an ensemble average since the small-scale irregularities in the density are assumed to be statistically homogeneous in the sky plane. See Rickett (1977, (2) ff) for a discussion of the assumptions that go into this definition of R_n . This is a useful function for describing small-scale structure in both the neutral and ionized media. In the ionized medium various scattering and scintillation processes can be observed, particularly toward compact background sources like pulsars, that show a very broad power-law spectrum of irregularities from scales of tens of parsecs down to a few earth radii (Armstrong et al. 1995). The power spectrum of the fluctuations in density is related to the correlation function by the autocorrelation theorem, illustrated on  Fig. 11-14. So the power spectrum for a two-dimensional measurement (i.e., at a single distance d) is

$$P_{2n}(\vec{q}) = \frac{1}{(2\pi)^2} \int d^2r R_n(\vec{r}) e^{-i\vec{q}\cdot\vec{r}} \quad (11.58)$$

where \vec{q} is a vector on the u, v plane, i.e., the conjugate space to \vec{r} on the plane of the sky. A similar three-dimensional Fourier transform defines P_{3n} based on $R_n(\vec{r})$ defined with the displacement, \vec{r} , as a three-dimensional offset vector in ( 11.57). This is needed for some kinds of observations that are sensitive to the integral of the density fluctuations over the line of sight, like time variations in the dispersion measure of a pulsar.

The spatial power spectrum of the H I fluctuations can be calculated easily from 21-cm data taken from an interferometer or aperture synthesis telescope, since each (u, v) spacing corresponds to a point on the two-dimensional Fourier transform of the sky brightness. Fluctuations in density become fluctuations in the brightness temperature of the line by ( 11.8), the interferometer measures the fringe visibility function (upper right box on  Fig. 11-14, illustrated

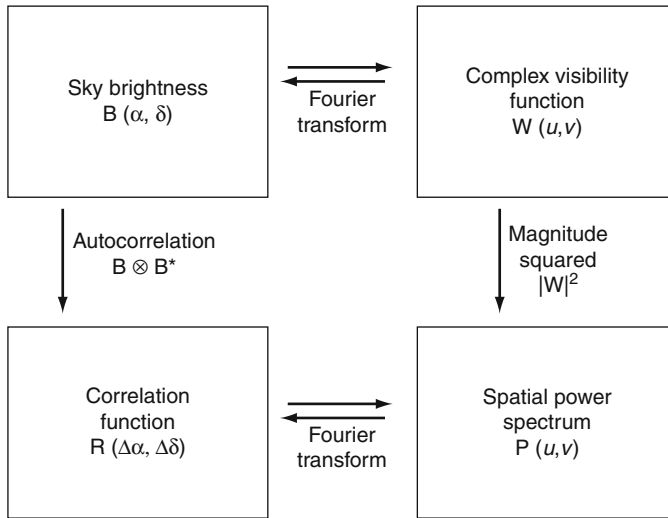


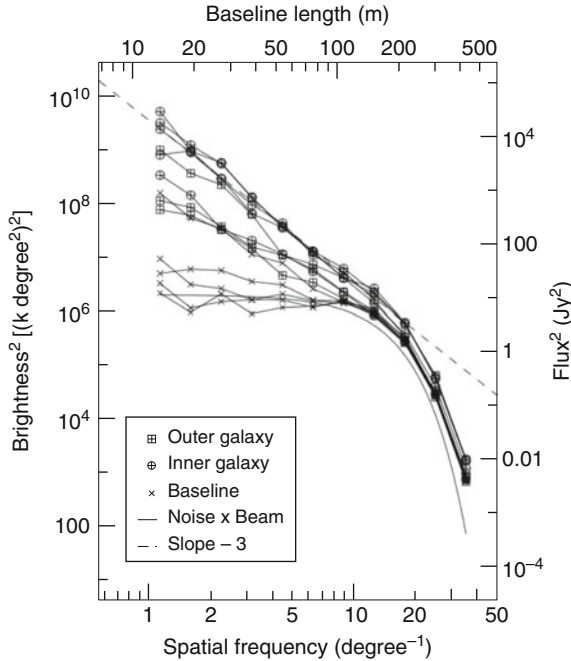
Fig. 11-14

The relationship between fluctuations in the density field, δn , their autocorrelation, R_n , and the Fourier transforms of both functions. These can be two dimensional, P_{2n} or three-dimensional functions, P_{3n} , depending on the measurement and the assumptions about the random process that generates δn , e.g., whether or not the fluctuations are isotropic. Figure 11-2 shows examples of the two boxes on the top row

in the middle column on Figure 11-2), and taking the magnitude squared of the fringe visibility function gives the spatial power spectrum averaged over the line-of-sight distance corresponding to the velocity range of the observation. When this is done for a narrow velocity width, Δv_r , at low latitude where the velocity gradient $\frac{dv}{ds}$ is large so that the line-of-sight depth, Δs , sampled in the density slice is small (11.41), then the spatial power spectrum is a good approximation of P_{2n} . Typically the slope of the P_{2n} function on a log-log scale is -2.5 to -3 (Crovisier and Dickey 1983; Green 1993). An example is shown in Figure 11-15. In some cases, averaging over a range of spectral channels that increases the depth of the sample, ΔS , to roughly a kpc changes the slope from -3 to -4 (Dickey et al. 2001). A steepening of the logarithmic slope by one unit is predicted by theories of turbulence (Lazarian and Pogosyan 2000), but this result is simply due to the change from a two-dimensional to a three-dimensional average in (11.58) and (11.57), and it does not establish any particular physical model of the turbulent process. The dynamics of the ISM on small scales are very different from the idealized Kolmogoroff model of turbulence in an incompressible medium (Kolmogorov 1941), and it is still controversial whether the structure of the ISM reflects a similar kinetic energy cascade from large to small scales.

7.2 Tiny Scale Structure

The current generation of aperture synthesis telescopes is limited in sensitivity to measuring structure in H I emission on angles of a few tens of arc-seconds or larger, but the high brightness temperature of some continuum sources makes it possible to measure the structure in the



■ Fig. 11-15

The spatial power spectrum of the H I emission at low latitudes (from Dickey et al. 2001). Channels with no line emission are marked with a *small x* and shown for reference. The interferometer beam causes the drop at high spatial frequencies, where the structure is unresolved. Logarithmic slope of -3 to -4 is very common for the spatial power spectrum of Galactic H I emission

absorption at much smaller angles, down to milli-arc seconds, using the technique of very long baseline interferometry (VLBI, Lazio et al. 2009). In the directions of several bright sources there are dramatic changes in optical depth, τ , over tiny scales as small as a few tens of astronomical units (AU). Similar structure in τ is indicated by observed time variations in absorption spectra toward pulsars (reviewed by Stanimirović et al. 2003). In many cases this structure can be explained as overlapping variations at different distances along the line of sight, with amplitudes consistent with the spatial power spectra of irregularities in n and T_c as well as the structure of the continuum source itself (Deshpande 2000), but in a few cases the τ variations are much stronger than elsewhere. If this τ variation is caused by density changes, the pressure fluctuations implied are huge (Heiles 1997), $\frac{P}{k} \simeq 10^5 \text{ K cm}^{-3}$ or more. This pressure would be a hundred times the mean ISM pressure or more. Indeed, there is good evidence from C I absorption that the distribution of pressure is very wide, and even bimodal, with a range in $\frac{P}{k}$ from less than 10^2 to more than $10^7 \text{ cm}^{-3} \text{ K}$ (Jenkins and Tripp 2011).

The variations in absorption seen on sub-arc-second scales do not give much information on the geometry of the structures involved. In particular, there is no reason to assume the absorption comes from tiny cloudlets, similar to but much smaller than the cloudlets seen in emission in the Lockman layer (► Fig. 11-3). Even for much larger structures seen in absorption there is a geometry problem, in that the column densities of cool gas implied by the optical depth

and spin temperature are much less than the density times the length scale seen in the plane of the sky. Thus a cloud with $\tau = 1$ over $\Delta v = 3 \text{ km s}^{-1}$ and $T_c = 60 \text{ K}$ must have column density $N \simeq 3 \cdot 10^{20} \text{ cm}^{-2}$; for density $n = 10^2 \text{ cm}^{-3}$ this implies a line-of-sight path length of $S \simeq 1 \text{ pc}$ (☛ 11.19). But absorbing clouds and cloud complexes typically have sizes in the plane of the sky of tens or even hundreds of pc. It may be that the shapes of the absorbing clouds are thin sheets usually seen face-on, or filaments usually seen side-on. But another possibility is that the CNM structure is like the network of walls in a sponge, with $N \ll nS$ because the path length, S , along any line of sight through the cloud, is mostly through a WNM substrate, like the water filling the cavities of a sponge. The very low optical depth clouds seen at high latitudes by Braun and Kanekar (2005) may be isolated examples of the building blocks of this CNM network.

8 Looking Ahead

There are many new observatories planned or under construction that will make great advances in the study of the Milky Way disk, but for the 21-cm line the Square Kilometer Array (SKA) is the one with the most promise. For the design and proof of concept of this ambitious facility, several pathfinder instruments are in development that will themselves allow an order of magnitude improvement over existing data in various combinations of sensitivity and resolution in the H I line. The Australian Square Kilometer Array Pathfinder (ASKAP, Johnston et al. 2007), the South African Karoo Array Telescope (MeerKAT, Jonas 2009), the Allen Telescope Array (ATA, Deboer et al. 2004), and the WSRT focal plane array upgrade (APERTIF, Verheijen et al. 2008) are all variations on a design for the SKA with a large number of small diameter antennas (LNSD). As these and other promising new telescopes become available for study of the Galactic disk in the 21-cm line it will be possible to undertake much more ambitious surveys of emission and absorption.

The recent installation of the Arecibo L-band Focal Plane Array (ALFA), a multi-beam receiver on the 305 m diameter Arecibo telescope, has made this extremely large aperture seven times faster than before for all kinds of surveys at $f \sim 1.4 \text{ GHz}$. Although neither MeerKAT, ASKAP, APERTIF, nor the ATA will have nearly as much collecting area as the Arecibo telescope, their survey speeds, $\frac{d\Omega}{dt}$, in square degrees per hour, will be faster at a given sensitivity because their fields of view are much larger. As aperture synthesis arrays, they can obtain much higher resolution than any single dish, but for a given total collecting area and number of receivers, the survey speed at a given brightness temperature threshold increases with finer angular resolution, θ , as $\frac{d\Omega}{dt} \propto \theta^{-4}$. The new telescopes offer the capability of mapping the H I emission with resolution of $10''\text{--}3'$, and brightness sensitivity $\sigma_T \ll 1 \text{ K}$. Proposals for these new instruments often showcase surveys of extragalactic continuum and line emission as their key science goals, but the astrophysical significance of surveys of the Galactic plane 21-cm emission and absorption is also a strong justification for the SKA and its precursors.

Surveys with the new facilities already under construction will provide maps of the H I emission either with a factor of 10 lower noise at the same resolution as the CGPS and SGPS ($\sigma_T \sim 0.1$ vs. 1 K), or with the same sensitivity but a factor of five improvement in resolution ($15''\text{--}20''$ vs. $1'\text{--}2'$), or any combination of these parameters. In addition, the number of absorption spectra toward extragalactic background sources will increase by a factor of 5–10 at

any given optical depth sensitivity, σ_τ . The combination of these advances will give comprehensive and detailed results for many functions and quantities that are only known roughly today, such as the scale height of the H I gas as a function of temperature, or the velocity distribution of clouds as a function of mass.

In addition to improving measurements of quantities that are already known roughly, new surveys can take on problems that are not tractable with current data. Some of these may be as follows:

- The speed and efficiency of condensation of the ISM from WNM to CNM to molecular clouds, and the survival of turbulent motions in the gas during these transitions. Comprehensive data on 21-cm absorption will show the spatial relationship between the CNM and the diffuse molecular gas, and give a much better tracing of the small-scale structure in the CNM down to tiny scales.
- The role of the magnetic field in the dynamics of the neutral atomic ISM and in establishing the balance between heating and cooling. This can be addressed through direct measurement of the field strength in the H I by the Zeeman splitting of the 21-cm line, and indirectly by observing Faraday rotation and linear polarization of the diffuse synchrotron emission of the Galactic cosmic ray population.
- Mass exchange between the disk and halo, including disk gas that flows up through interstellar chimneys like that in [Fig. 11-13](#) and then falls back to the disk as a “Galactic Fountain” (Bregman 1980), infall of gas accreted from outside the Milky Way like the Magellanic Stream, and even the possibility of a wind from the disk that leaves the Galaxy.
- The astrophysical conditions that regulate star formation. The variations in gas properties with R and z will show the factors that control the flow of mass into gravitationally unstable molecular clouds. In turn, galaxy evolution depends on the local pressure, radiation field, and abundance of heavy elements and dust grains, as well as the more basic surface density of gas and the Toomre Q stability parameter (Kennicutt 1989).

As the history of the universe becomes better understood, numerical models of galaxy formation and evolution will need to include more sophisticated treatment of the astrophysics of the ISM. Only Galactic astronomy can provide the astrophysical detail needed to make such numerical models realistic on scales smaller than a few kpc. The cycle shown in [Fig. 11-1](#) is well established on the stellar evolution (left) side, but the flow of matter through the warm and cool ISM phases is not so well understood, even in the Milky Way. The diversity of galaxy types and star formation histories shows how much variation there can be in galaxy evolution. Understanding of the causes of this diversity may come by detailed observations of the ISM of the Milky Way as well or better than by surveys of distant galaxies. Galactic astronomers have good reason to say, “Galaxy evolution begins at home.”

Cross-References

- [Astrophysics of Galactic Charged Cosmic Rays](#)
- [Dynamics of Disks and Warps](#)
- [High-Velocity Clouds](#)
- [History of Dark Matter in Galaxies](#)
- [Galactic Distance Scales](#)

- [Gamma-Ray Emission of Supernova Remnants and the Origin of Galactic Cosmic Rays](#)
- [Magnetic Fields in Galaxies](#)
- [Mass Distribution and Rotation Curve in the Galaxy](#)

References

- Anderson, L. D., & Bania, T. M. 2009, *ApJ*, 690, 706
- Armstrong, J. W., Rickett, B. J., & Spangler, S. R. 1995, *ApJ*, 443, 209
- Ballesteros-Paredes, J., & Mac Low, M.-M. 2002, *ApJ*, 570, 734
- Belfort, P., & Crovisier, J. 1984, *A&A*, 136, 368
- Binney, J., & Merrifield, M. R. 1998, *Galactic Astronomy* (Princeton: Princeton University Press)
- Brand, J., & Blitz, L. 1993, *A&A*, 275, 67
- Brandt, J. C. 1960, *ApJ*, 131, 293
- Braun, R., & Kanekar, N. 2005, *A&A*, 436, L53
- Bregman, J. N. 1980, *ApJ*, 236, 577
- Burton, W. B. 1988, in *Galactic and Extragalactic Radio Astronomy*, eds. G. L. Verschuur, & K. Kellerman (2nd ed.; New York: Springer-Verlag), 295–358
- Burton, W. B., & Liszt, H. S. 1993, *A&A*, 274, 765
- Celnick, W., Rohlfis, K., & Braunsfurth, E. 1979, *A&A*, 76, 24
- Chandrasekhar, S., & Munch, G. 1952, *ApJ*, 115, 103
- Clark, B. G. 1965, *ApJ*, 142, 1398
- Condon, J. J. 1989, in *Galactic and Extragalactic Radio Astronomy*, eds. G. L. Verschuur, & K. Kellerman (2nd ed.; New York: Springer-Verlag), 522
- Crovisier, J., & Dickey, J. M. 1983, *A&A*, 122, 282
- Dalgarno, A., & McCray, R. A. 1972, *ARAA*, 10, 375
- de Avillez, M. A., & Breitschwerdt, D. 2007, *ApJ Lett*, 665, L35
- de Blok, W. J. G., Walter, F., Brinks, E., Trachternach, C., Oh, S.-H., & Kennicutt, R. C. 2008, *AJ*, 136, 2648
- Deboer, D., Welch, W. J., Dreher, J., Tarter, J., Blitz, L., et al. 2004, *Proc SPIE*, 5489, 1021
- Deshpande A. 2000, *MNRAS*, 317, 199
- Dickey, J. M., & Lockman, F. J. 1990, *Ann Rev Astron Astrophys*, 28, 215
- Dickey, J. M., McClure-Griffiths, N. M., Stanimirović, S., Gaensler, B. M., & Green, A. J. 2001, *ApJ*, 561, 264
- Dickey, J. M., McClure-Griffiths, N., Gaensler, B., & Green, A. 2003, *ApJ*, 585, 801
- Dickey, J. M., Strasser, S., Gaensler, B. M., Haverkorn, M., Kavars, D., McClure-Griffiths, N. M., Stil, J., & Taylor, A. R. 2009, *ApJ*, 693, 1250
- Draine, B. T. 1978, *ApJ Suppl*, 36, 595
- Draine, B. T. 2011, *Physics of the Interstellar and Intergalactic Medium*, Princeton: Princeton University Press
- Dwarakanath, K. S., Carilli, C. L., & Goss, W. M. 2002, *ApJ*, 567, 940
- Elmegreen, B. G., & Scalzo, J. 2004, *ARA&A*, 42, 211
- Ferrière, K. 2001, *Rev Mod Phys* 73, 1031.
- Ferrière, K. M., Zweibel, E. G., & Shull, J. M. 1988, *ApJ*, 332, 984
- Ferrière, K., Gillard, W., & Jean, P. 2007, *A&A*, 467, 611
- Field, G. B. 1959, *ApJ*, 129, 536
- Field, G. B., & Saslaw, W. C. 1965, *ApJ*, 142, 568
- Field, G. B., Goldsmith, D. W., & Habing, H. J. 1969, *ApJ Lett*, 155, L149
- Ford, H. A., McClure-Griffiths, N. M., Lockman, F. J., Bailin, J., Calabretta, M. R., Kalberla, P. M. W., Murphy, T., & Pisano, D. J. 2008, *ApJ*, 688, 290
- Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, *Phys Rep*, 433, 181–301
- Gibson, S. J., Taylor, A. R., Higgs, L. A., Brunt, C. M., & Dewdney, P. E. 2005a, *ApJ*, 626, 195
- Gibson, S. J., Taylor, A. R., Stil, J. M., Dewdney, P. E., Brunt, C. M., & Strasser, S. T. 2005b, *JRASC*, 99, 129
- Gould, R. J. 1994, *ApJ*, 423, 522
- Green, D. A. 1993, *MNRAS*, 262, 327
- Heiles, C. 1997, *ApJ*, 481, 193
- Heiles, C., & Troland, T. H. 2003a, *ApJ Suppl*, 145, 329
- Heiles, C., & Troland, T. H. 2003b, *ApJ*, 586, 1067
- Heiles, C., & Troland, T. H. 2005, *ApJ*, 624, 773
- Hollenbach, D. J., & Tielens, A. G. G. M. 1997, *ARAA*, 35, 179
- Holmberg, J., & Flynn, C. 2004, *MNRAS*, 352, 440
- Howard, C. D., Rich, R. M., Clarkson, W., Mallery, R., Kormendy, J., et al. 2009, *ApJ*, 702, 153
- Jenkins, E. B., & Tripp, T. M. 2011, *ApJ*, 734, 65
- Johnston, S., Bailes, M., Bartel, N., Baugh, C., Bietenholz, M., et al. 2007, *PASA*, 24, 174
- Jonas, J. L. 2009, *Proc IEEE*, 97, 1522
- Kalberla, P. M. W., Burton, W. B., Hartmann, D., Arnal, E. M., Bajaja, E., Morras, R., & Pöppel, W. G. L. 2005, *A&A*, 440, 775
- Kalberla, P. M. W., Dedes, L., Kerp, J., & Haud, U. 2007, *A&A*, 469, 511

- Kalberla, P. M. W., & Dedes, L. 2008, *A&A*, 487, 951
- Kalberla, P. M. W., & Kerp, J. 2009, *Ann Rev Astron Astrophys*, 47, 27
- Kalberla, P. M. W., McClure-Griffiths, N. M., Pisano, D. J., Calabretta, M. R., & Ford, H.A., et al. 2010, *A&A* 521, 17
- Kang, J., & Koo, B.-C. 2007, *ApJ Suppl*, 173, 85
- Kavars, D. W., Dickey, J. M., McClure-Griffiths, N. M., Gaensler, B. M., & Green, A. J. 2005, *ApJ*, 626, 887
- Kennicutt, R. C. 1989, *AJ*, 344, 685
- Kolmogorov, A. 1941, *Dokl Akad Nauk SSSR*, 30, 301
- Kolpak, M. A., Jackson, J. M., Bania, T. M., Clemens, D. P., & Dickey, J. M., 2003, *ApJ*, 582, 756
- Krco, M., Goldsmith, P. F., Brown, R. L., & Li, D. 2008, *ApJ*, 689, 276
- Kuijken, K., & Gilmore, G. 1989, *MNRAS*, 239, 651
- Lazarian, A., & Pogosyan, D. 2000, *ApJ*, 537, 720
- Lazio, T. J. W., Brogan, C. L., Goss, W. M., & Stanimirović, S. 2009, *AJ*, 137, 4526
- Levine, E. S., Heiles, C., & Blitz, L. 2006, *ApJ*, 643, 881
- Levine, E. S., Heiles, C., & Blitz, L. 2008, *ApJ*, 679, 1288
- Liszt, H. S. 1983, *ApJ*, 275, 163
- Lockman, F. J., & Gehman, C. 1991, *ApJ*, 382, 182
- Lockman, F. J., Hobbs, L. M., & Shull, J. M. 1986, *ApJ*, 302, 432
- Malhotra, S. 1995, *ApJ*, 448, 138
- McClure-Griffiths, N. M., & Dickey, J. M. 2007, *ApJ*, 671, 427
- McClure-Griffiths, N. M., et al. 2003, *ApJ*, 594, 833
- McClure-Griffiths, N. M., Dickey, J. M., Gaensler, B. M., Green, A. J., Haverkorn, M., & Strasser, S. 2005, *ApJ Suppl*, 158, 178
- McClure-Griffiths, N. M., Pisano, D. J., Calabretta, M., Ford, H. A., Lockman, F. J., et al. 2009, *ApJ Suppl*, 181, 398
- McKee, C. F., & Ostriker, J. P. 1977, *ApJ*, 218, 148
- Mebold, U., Hachenberg, O., & Laury-Micoulaut, C. A. 1974, *A&A*, 30, 329
- Mebold, U., Dusterberg, C., Dickey, J. M., Staveley-Smith, L., Kalberla, P., Muller, H., & Osterberg, J. 1997, *ApJ Lett*, 490, 65
- Norman, C. A., & Ikeuchi, S. 1989, *ApJ*, 345, 372
- Oort, J. H. 1955, in *Vistas in Astronomy*, Vol. 1, ed. A. Beer, (London: Pergamon), 607
- Parker, E. N. 1966, *ApJ*, 145, 811
- Rickett, B. J. 1977, *ARAA*, 15, 479
- Rohlfs, K., & Wilson, T. L. 2006, *Tools of Radio Astronomy* (Berlin: Springer)
- Roman-Duval, J., Jackson, J. M., Heyer, M., Johnson, A., Rathborne, J., Shah, R., & Simon, R. 2009, *ApJ*, 699, 1153
- Rownd, B. K., Dickey, J. M., & Helou, G. 1994, *AJ*, 108, 1638
- Slavin, J. D., & Cox, D. P. 1992, *ApJ*, 392, 131
- Spitzer, L., Jr. 1977, *Physical Processes in the Interstellar Medium* (New York: Wiley)
- Stanimirović, S., & Heiles, C. 2005, *ApJ*, 631, 371
- Stanimirović, S., Staveley-Smith, L., Dickey, J. M., Sault, R. J., & Snowden, S. L. 1999, *MNRAS*, 302, 417
- Stanimirović, S., Weisberg, J. M., Hedden, A., Devine, K. E., & Green, J. T. 2003, *ApJ*, 598, 23
- Stanimirović, S., Putman, M., Heiles, C., Peek, J. E. G., Goldsmith, P. F., et al. 2006, *ApJ*, 653, 1210
- Stil, J. M., Taylor, A. R., Dickey, J. M., Kavars, D. W., Martin, P. G., et al. 2006, *AJ*, 132, 1158
- Strasser, S. T. 2006, Ph.D. Thesis, University of Minnesota, Chap. 4
- Strasser, S. T., & Taylor, A. R. 2004, *ApJ*, 603, 560
- Strasser, S. T., Dickey, J. M., Taylor, A. R., Boothroyd, A. I., Gaensler, B. M., et al. 2007, *AJ*, 134, 2252
- Tamburro, D., Rix, H.-W., Leroy, A. K., Mac Low, M.-M., Walter, F., et al. 2009, *ApJ*, 137, 4424–4435
- Taylor, G. B., Carilli, C. L., & Perley, R. A., eds. 1999, *Synthesis Imaging in Radio Astronomy II* (San Francisco: ASP)
- Taylor, A. R., Gibson, S. J., Peracaula, M., Martin, P. G., Landecker, T. L., et al. 2003 *AJ*, 125, 3145
- Thompson, A. R., Moran, J. M., & Swenson, G. W. 1986, *Interferometry and Aperture Synthesis in Radio Astronomy* (New York: Wiley)
- van de Hulst, H. C., Muller, C. A., & Oort, J. H. 1954, *BAIN*, 12, 117–149
- Vázquez-Semadeni, E., Passot, T., & Pouquet, A. 1995, *ApJ*, 441, 702
- Verheijen, M. A. W., Oosterloo, T. A., van Cappellen, W. A., Bakker, L., Ivashina, M. V., & van der Hulst, J. M. 2008, *AIPC*, 1035, 265
- Weiner, B. J., & Sellwood, J. A. 1999, *ApJ*, 524, 112
- Wolfire, M. G., Hollenbach, D., McKee, C. F., Tielens, A. G. G. M., & Bakes, E. L. O. 1995, *ApJ*, 443, 152
- Wolfire, M. G., McKee, C. F., Hollenbach, D., & Tielens, A. G. G. M. 2003, *ApJ*, 587, 278

12 High-Velocity Clouds

Bart P. Wakker¹ · Hugo van Woerden²

¹Supported by NASA and NSF; affiliated with Department of Astronomy, University of Wisconsin, Madison, WI, USA

²Kapteyn Astronomical Institute, Rijksuniversiteit Groningen, Groningen, The Netherlands

1	<i>Introduction</i>	589
2	<i>Sky Maps: Clouds and Complexes</i>	591
2.1	The Deviation Velocity	591
2.2	Mapping the HVCs	592
2.3	Features of the HVC Sky	595
2.4	Sky Coverage	595
2.5	HVC Kinematics	597
2.6	Ionized HVCs	599
3	<i>The Determination of Cloud Distances and Metallicities</i>	600
3.1	HVC Distances from Indirect Methods	600
3.1.1	H α Intensity	600
3.1.2	Pressure Equilibrium	601
3.1.3	Velocity Gradients	601
3.1.4	The Virial Theorem	601
3.2	Absorption-Line Method for Determining Distances	602
3.2.1	Suitable Stellar Candidates Projected onto the HVC	603
3.2.2	Spectra and Photometry to Characterize the Candidates	603
3.2.3	High-resolution Spectra of Stars at Distances Bracketing that of the HVC	603
3.3	Considerations for Metallicity Measurements	604
3.3.1	HI Small-Scale Structure	605
3.3.2	Depletion onto Dust	605
3.3.3	Presence of Ionized Hydrogen	605
3.3.4	Elemental Ionization Fractions	605
3.4	Strength of the Absorption Lines	606
3.5	Measured Distances and Metallicities	609
4	<i>Physical Properties of the HVCs</i>	610
4.1	Small-Scale Structure	610
4.2	Timescales	613
4.2.1	The Time it Will Take for a Cloud to Reach the Galactic Plane	613
4.2.2	The Time for the Cores to Shift Substantially Relative to Each Other	613
4.2.3	The Time a Core takes to Move Across its Own Width	614
4.2.4	The Time for a Core to Double its Size	614

4.3	Ionization Structure and Volume Density	614
4.4	Molecules and Dust	615
5	<i>Hot Gas Associated with HVCs</i>	617
6	<i>HVCs in and Around Other Galaxies</i>	622
6.1	A Face-on Galaxy: M 101	622
6.2	An Edge-on Galaxy: NGC 891	623
6.3	A High-Inclination Galaxy: NGC 2403	623
6.4	A Low-Inclination Galaxy: NGC 6946	624
6.5	The Nearest Spiral Galaxies: Messier 31 and 33	625
6.6	High-Velocity and Extraplanar Gas in Other Galaxies	626
6.7	Accretion of Gas by Galaxies	627
7	<i>Origins of the High-Velocity Clouds</i>	627
7.1	The Galactic Fountain	628
7.2	Tidal Streams	630
7.3	Low-Metallicity Accretion	631
7.3.1	Fragmentation in a Hot Halo	632
7.3.2	Cold, Filamentary Streams	633
7.3.3	Sweeping up of Coronal Gas by Fountain Flows	634
7.4	Summary of Origins	634
	<i>References</i>	636

Abstract: The high-velocity clouds (HVCs) are gaseous objects that do not partake in differential galactic rotation, but instead have anomalous velocities. They trace energetic processes on the interface between the interstellar material in the Galactic disk and intergalactic space. Three different processes appear to be responsible for the formation of HVCs. First, supernovae in the Galactic disk create hot gas that vents into the halo, cools and rains back down, in a process generically termed the “Galactic Fountain,” in which gas circulates between the disk and halo at a rate of a few $M_{\odot} \text{ yr}^{-1}$. This implies that the interstellar medium (ISM) circulates through the halo on timescales of a Gyr. Second, gas streams are formed by tides working on nearby dwarf galaxies (with a possible contribution from ram pressure); this applies specifically to the Magellanic Stream, which was extracted from the Small Magellanic Cloud. Third, low-metallicity clouds are accreting onto the Milky Way, at a present-day rate of about $0.4 M_{\odot} \text{ yr}^{-1}$. Such infall causes the Milky Way to grow and continue forming stars. The source of the infalling material may lie in the cooling of hot ($T > 10^6 \text{ K}$) intergalactic gas that permeates space, or in cold ($T < 10^5 \text{ K}$) accretion streams that are theoretically predicted to transport material from intergalactic filaments to galaxies. This chapter describes the observed locations, velocities, and physical conditions of the HVCs. Also included is a discussion of the methods used to derive their distances and metallicities, as well as of the resulting values. Finally, the different origins of the HVCs are discussed.

1 Introduction

Spiral galaxies have a number of different constituents. Dark matter is responsible for most of the gravity well, stars emit light, and everything is permeated by the interstellar medium (ISM), from which stars form. In turn, the structure of the ISM is determined by the feedback of matter and energy from the stars, as well as by the infall of new material.

Most of the dense ISM is concentrated in the Galactic plane, and rotates around the Galactic Center, just as the stars do. However, several energetic processes lead to gas moving at anomalous velocities, and this gas plays a role in our understanding of the evolution of the Galaxy. These processes include the “Galactic Fountain,” caused by supernovae that heat the ISM and lift it up a few kpc, into the lower halo, where it cools and rains back down (Shapiro and Field 1976; Bregman 1980; Kahn 1981; de Avillez and Breitschwerdt 2005). This circulation redistributes heavy elements and energy; the rate of circulation is likely to be dependent on the supernova rate, while the radial extent of the mixing will depend on the precise balance between various factors, such as the supernova rate, the gravitational potential, and the thermal evolution of the gas. A second process that generates gas at anomalous velocities is the infall of low-metallicity gas, which provides new fuel for star formation. Such gas can be provided by interactions with passing galaxies (i.e., as tidal streams – see e.g., Gardiner and Noguchi 1996; Mastrogiuseppe et al. 2005; Connors et al. 2006; Besla et al. 2007), or by instabilities in the massive, large (>200 kpc), hot (10^6 K) coronae of ionized gas that surround the galaxies (Maller and Bullock 2004; Stocke et al. 2006; Wakker and Savage 2009). Alternatively, cold ($T < 10^5 \text{ K}$) gas may stream into galaxies along intergalactic filaments (Kereš and Hernquist 2009), or it may be swept along by Galactic Fountain clouds (Fraternali and Binney 2008; Marasco et al. 2011). Such newly accreted gas is needed in models of the chemical evolution of the Galaxy, which require star formation to be fed by continuing accretion, at a present-day inflow rate of $\sim 0.4 M_{\odot} \text{ yr}^{-1}$ of material with a metallicity $Z \sim 0.1$ times solar (Chiappini 2008).

Observationally, the gas moving at anomalous velocities is seen in the form of the “high-velocity clouds” (HVCs). Over time, this term has undergone a gradual change in meaning. A detailed summary of the historical development of the study of the HVCs was presented by van Woerden et al. (2004) in Chapter 1 of their monograph on the high-velocity clouds. Originally, the term HVC was applied to interstellar absorption lines at velocities $>20 \text{ km s}^{-1}$ relative to the Sun, which were seen in the spectra of high-latitude stars (Adams 1949; Münch 1952; Schlüter et al. 1953). Later, it was mostly applied to high-latitude neutral hydrogen clouds seen in 21-cm emission with velocities relative to the Local Standard of Rest (LSR) $>80 \text{ km s}^{-1}$ (Muller et al. 1963; Oort 1966). HI clouds at velocities $|v_{\text{LSR}}| = 40\text{--}80 \text{ km s}^{-1}$ were called “intermediate-velocity clouds” (IVCs). Throughout the 1970s and 1980s, 90 and 100 km s^{-1} were also (inconsistently) considered as velocity limits. Wakker (1991) adapted the definition by proposing to use the “deviation velocity” (v_{DEV}), the difference between the observed velocity of the gas and the maximum velocity that can be understood in a simple model of differential galactic rotation. In this definition HVCs have $|v_{\text{DEV}}| > 90 \text{ km s}^{-1}$ and IVCs have $|v_{\text{DEV}}| = 30\text{--}90 \text{ km s}^{-1}$. This has been the working definition since. However, it has not always been strictly adhered to, as the current HVC and IVC catalogues are still based on the old definitions.

Following their discovery, research on HVCs went into three main directions. First, the mapping of the HVC sky, which culminated in the surveys of Giovanelli (1980), Bajaja et al. (1985), and Hulsbosch and Wakker (1988), and in the whole-sky HVC catalogue and definition of HVC complexes (groups of clouds with similar location and velocity) by Wakker and van Woerden (1991). Second, detailed mapping and characterizing of the physical properties of individual clouds (Giovanelli et al. 1973; Davies et al. 1976; Giovanelli and Haynes 1977; Schwarz and Oort 1981; Wakker and Schwarz 1991). Third, attempts at finding an explanation for their origin, with Oort (1966, 1970) presenting the first comprehensive list. His discussion is still generally valid, although the details were much expanded and refined over the following 40 years. Major refinements include the idea of the Galactic Fountain (put forward by Shapiro and Field 1976 and applied to HVCs by Bregman 1980), the understanding of the Magellanic Stream as a tidal feature (proposed by Fujimoto and Sofue 1976 and Lin and Lynden-Bell 1982), and the arguments that HVCs are distant, possibly even Local Group objects (Blitz et al. 1999; Braun and Burton 1999).

Starting in the 1980s, but taking off in the late 1990s, developments in instrumentation and the availability of space observatories and large ground-based telescopes led to the detection of absorption associated with HVCs in the spectra of distant halo stars and UV-bright extragalactic objects (i.e., QSOs and Seyfert galaxies). These studies yielded measurements of HVC distances and metallicities (e.g., Kuntz and Danly 1996; Lu et al. 1998; van Woerden et al. 1999a; Wakker et al. 1999, 2007, 2008; Richter et al. 2001b). When UV absorption-line studies of tracers of hot gas (O VI, C IV) in directions away from the 21-cm HVCs also started showing high-velocity gas (Sembach et al. 1995, 2003; Fox et al. 2004, 2006; Lehner and Howk 2011) the term HVC was extended beyond just the HI clouds. In addition, deep 21-cm observations of other galaxies have revealed HI gas moving at anomalous velocities (van der Hulst and Sancisi 1988; Braun and Thilker 2004; Fraternali et al. 2004, Oosterloo et al. 2007), and by analogy the HVC name was also applied to such objects.

This chapter will use an expansive definition of the term “HVC,” taking it to include the “classical” 21-cm HVCs and IVCs, as well as the high-velocity gas seen in UV absorption lines and the extragalactic anomalous-velocity clouds. ➤ Section 2 presents the observational definition for the Galactic HVCs and discusses the objects that are seen in the sky. ➤ Section 3 summarizes the methods used to determine the distances and metallicities of the clouds, with

our current knowledge of these quantities given in [Sect. 3.5](#). Using these results, [Sect. 4](#) summarizes the physical properties of the clouds. The UV absorption line studies of the hot component of the high-velocity cloud phenomenon are discussed in [Sect. 5](#). Extragalactic HVCs are summarized in [Sect. 6](#). Finally, in [Sect. 7](#) the different explanations that have been put forward for the origin of the HVCs are evaluated.

2 Sky Maps: Clouds and Complexes

2.1 The Deviation Velocity

To best capture the idea that the HVCs represent gas that does not take part in galactic rotation, a basic observational definition was proposed by Wakker (1991). He defined the “deviation velocity,” v_{DEV} , which is the difference between the observed velocity and the maximum velocity that can easily be understood in terms of differential galactic rotation. For a particular direction it is found by calculating

$$\begin{aligned} v_{\text{DEV}} &= v_{\text{LSR}} - v_{g,\text{min}} \text{ if } v_{\text{LSR}} < 0 \\ v_{\text{DEV}} &= v_{\text{LSR}} - v_{g,\text{max}} \text{ if } v_{\text{LSR}} > 0. \end{aligned}$$

Here v_{LSR} is the observed velocity relative to the Local Standard of Rest (LSR) and $v_{g,\text{min,max}}$ are the minimum and maximum possible velocities for rotating disk gas in this direction, $v_g(l, b, d)$. The latter can be found using geometrical relations giving the galactocentric radius (R) and height above the plane (z) as function of distance in the line of sight (d) for a given longitude (l) and latitude (b), combined with a prescription for $v(R)$, the rotation velocity as function of galactocentric radius (i.e., the Galactic rotation curve). It also requires a value for R_0 , the distance of the Sun to the Galactic Center, which is about 7.9 kpc (see [Chap. 16](#) by Feast in this volume).

$$\begin{aligned} v_g(l, b, d) &= \left(\frac{R_0}{R} v(R) - v(R_0) \right) \sin l \cos b \\ R(l, b, d) &= R_0 \sqrt{\cos^2 b \left(\frac{d}{R_0} \right)^2 - 2 \cos b \cos l \left(\frac{d}{R_0} \right) + 1} \\ z(l, b, d) &= d \sin b. \end{aligned}$$

For this chapter, the Galactic rotation curve is assumed to be flat with velocity $v(R) = 220 \text{ km s}^{-1}$ at radii $R > 0.5 \text{ kpc}$, and solid-body closer to the center. Then these three relations can be used to calculate v_{LSR} for all distances between 0 and d_{max} , with d_{max} defined as the distance where the sightline leaves the disk. Many different prescriptions are possible for the edge of the disk. The one that is used in this chapter assumes that the disk has a radius of 26 kpc and thickness $z_1 = 2 \text{ kpc}$ for $R < R_0$, flaring parabolically to $z_2 = 6 \text{ kpc}$ at $R = 3R_0$, so that:

$$\begin{aligned} R_{\text{max}} &= 26 \text{ kpc} \\ z_{\text{max}} &= z_1 \text{ if } R < R_0 \\ z_{\text{max}} &= z_1 + (z_2 - z_1) \times \frac{(R/R_0 - 1)^2}{4} \text{ if } R > R_0. \end{aligned}$$

2.2 Mapping the HVCs

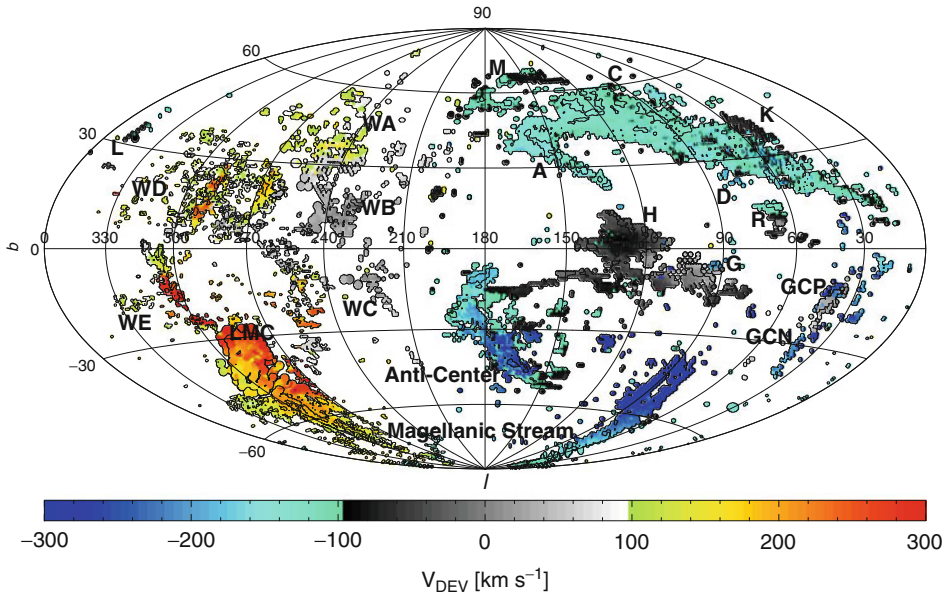
The history of the mapping of the HVCs was reviewed by Wakker and van Woerden (1997) and Wakker (2004). Most of the early observations covered only part of the sky or had rough sampling relative to the size of the telescope beam. The first all-sky survey of high-velocity gas with $|\nu_{\text{LSR}}| > 90 \text{ km s}^{-1}$ was provided by the combination of the lists of Hulsbosch and Wakker (1988) and Bajaja et al. (1985). The former was made using the Dwingeloo telescope and covered the northern sky (declinations $> -23^\circ$) on a $1^\circ \times 1^\circ$ grid, with a 0.5° beam, 16 km s^{-1} velocity resolution, and $5\text{-}\sigma$ detection limit 0.05 K , while the latter (with the Argentinian Villa Elisa telescope) covered declinations $< -18^\circ$ on a $2^\circ \times 2^\circ$ grid, with detection limit 0.08 K . Both datasets were published as a list of HVC profile components, giving longitude, latitude, LSR velocity, and peak brightness temperature. Information about line widths and shapes was mostly lost, but for a typical line width of 20 km s^{-1} , the 5σ detection limits correspond to column densities of ~ 2.0 and $3.0 \times 10^{18} \text{ cm}^{-2}$, respectively. Wakker and van Woerden (1991) used these two datasets to construct the first (and so-far only) all-sky catalogue of the HVCs.

A more complete survey of the HI sky, the “Leiden-Argentina-Bonn” or LAB survey was constructed by Kalberla et al. (2005). This survey combined the northern (declination $> -30^\circ$) “Leiden-Dwingeloo Survey” of Hartmann and Burton (1997) with a complementary survey of the southern sky, again made using the Villa Elisa telescope (Arnal et al. 2000). Both surveys cover the sky on a $0.5^\circ \times 0.5^\circ$ grid, with 1.03 km s^{-1} velocity channels, an rms noise of 0.07 K , and are fully (internally consistently) corrected for stray-radiation effects. Morras et al. (2000) used the southern survey to construct a much improved list of southern HVC components. Wakker et al. (2011) discovered that the published LAB spectra still require a small correction, in the sense that an underlying broad gaussian ($\nu = -22 \text{ km s}^{-1}$, $T_B = 0.0473 \text{ K}$, $\text{FWHM} = 167 \text{ km s}^{-1}$, corresponding to $N(\text{HI}) = 1.5 \times 10^{19} \text{ cm}^{-2}$) needs to be removed. This component is most likely either a residual baseline fitting error, or a small error in the stray radiation correction.

Compared to the combined Hulsbosch and Wakker (1988) plus Morras et al. (2000) survey, the LAB survey has advantages and disadvantages for the study of the HVCs. The LAB survey allows more detailed mapping (on a $0.5^\circ \times 0.5^\circ$ grid instead of a $1^\circ \times 1^\circ$ grid), has better velocity resolution (1.25 km s^{-1} vs 16 km s^{-1}), and it covers the IVCs. However, since the integration times for the LAB survey were much shorter than what was the case for the older surveys, the 5σ detection limit for a cloud with line width 20 km s^{-1} is about $3 \times 10^{18} \text{ cm}^{-2}$, slightly worse than in the Hulsbosch and Wakker (1988) survey. To achieve the same sensitivity in the LAB data therefore requires smoothing to a 1° beam. Thus, the ~ 150 small, faint ($< 1^\circ$ diameter; $T_{B,\text{peak}} < 0.08 \text{ K}$; $N(\text{HI}) < 3 \times 10^{18} \text{ cm}^{-2}$) clouds in the Hulsbosch and Wakker (1988) survey are below the LAB detection limit. Similarly, the faint edges of the large HVC complexes are more easily seen in the older survey. A combination of the old HVC surveys and the LAB data is therefore necessary to extract the most information.

🔍 *Figure 12-1* presents an all-sky map of the HVCs as seen in the combined Hulsbosch and Wakker (1988) and Morras et al. (2000) lists. Clouds in grey have $|\nu_{\text{LSR}}| > 90 \text{ km s}^{-1}$, but $|\nu_{\text{DEV}}| < 90 \text{ km s}^{-1}$, meaning that they should properly be classified as IVCs. 🔍 *Figures 12-2* and 🔍 *12-3* give a map of the IVCs at positive and negative velocities, based on the LAB survey. Detailed maps of many individual HVC and IVC clouds can be found in Wakker (2001) and Wakker et al. (2008).

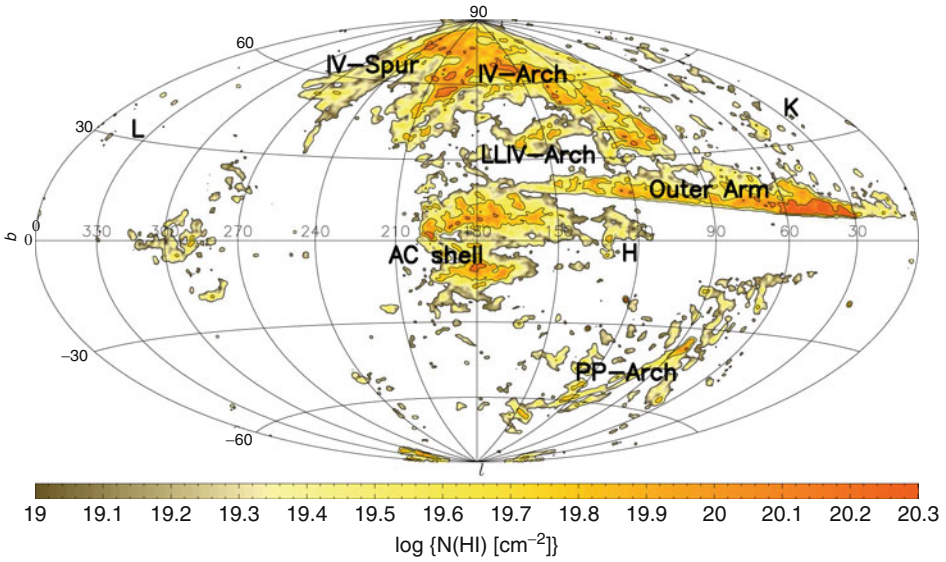
An improvement in the mapping of HVCs will be possible using some recent surveys. The Parkes GASS survey (Kalberla et al. 2010) provides data for $|\nu_{\text{LSR}}| < 468 \text{ km s}^{-1}$ at



■ Fig. 12-1

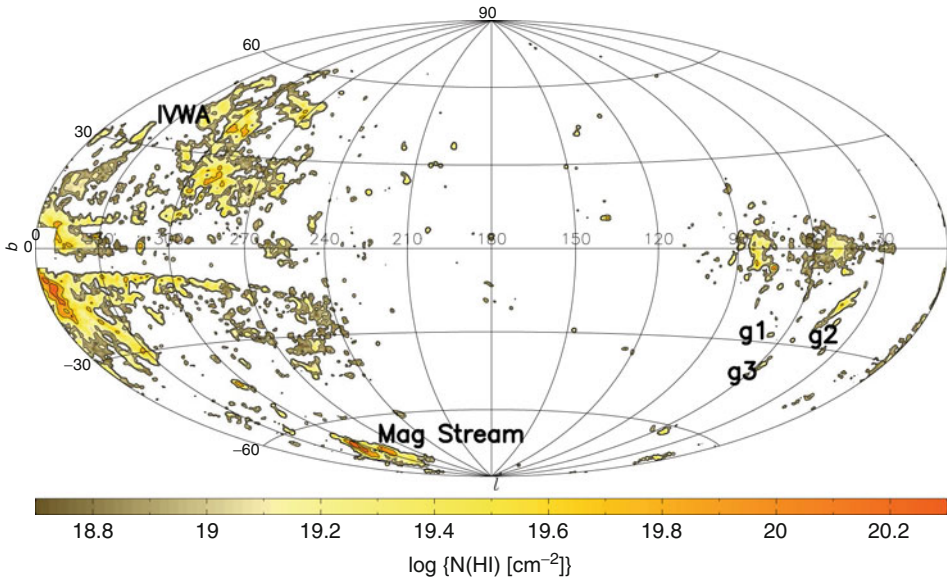
Aitoff projection all-sky map of the HVCs, in galactic coordinates, with the Anti-Center in the middle (Based on Hulsbosch and Wakker (1988) and Morras et al. (2000)). Contour levels are at brightness temperatures of 0.05 and 0.5 K. Colors code deviation velocities, with the scale given by the bar at the bottom. HVCs were selected using $|v_{\text{LSR}}| > 90 \text{ km s}^{-1}$. Gray clouds have $|v_{\text{LSR}}| > 90 \text{ km s}^{-1}$, but $|v_{\text{DEV}}| < 90 \text{ km s}^{-1}$. Labels give the names of the cloud complexes

declinations $< +1^\circ$ with a $14.4'$ beam and 0.057 K rms in a 1.0 km s^{-1} channel. The sky north of -5° declination will be covered by the Effelsberg HI survey (EBHIS; Winkel et al. 2010) on a $9.5'$ grid with 0.09 K rms per 1.25 km s^{-1} channel. The Arecibo GALFA-HI survey (Peek et al. 2011) covers the sky between declinations -1° and $+38^\circ$ with a $3.4'$ beam, 0.18 km s^{-1} channels, and rms 0.08 K in a 1 km s^{-1} channel. For HVCs with line widths 20 km s^{-1} , these surveys thus have single-beam 5σ detection limits of 2.5×10^{18} , 4.4×10^{18} , and $3.5 \times 10^{18} \text{ cm}^{-2}$, respectively. For a beam-filling cloud that is larger than $36'$ in size, smoothing to a $36'$ equivalent beam can potentially reduce this further to 1.0×10^{18} , 1.1×10^{18} , and $0.35 \times 10^{18} \text{ cm}^{-2}$, respectively. However, most faint clouds are small and/or have structure, so that in practice the faintest detections will have column densities larger than these numbers. Thus, compared to the LAB data, these surveys will allow making maps with increased angular resolution (by factors of 2.5, 3.8, and 10.6, respectively). However, for mapping the extended outer parts of the HVCs and searching for faint clouds, these new surveys are not much more sensitive than the combined Hulsbosch and Wakker (1988) and Morras et al. (2000) datasets, even after smoothing to the equivalent 16 km s^{-1} and $36'$ resolution.



■ Fig. 12-2

Aitoff projection all-sky map of the column density of the intermediate-velocity gas with deviation velocity between -90 and -35 km s^{-1} , based on the LAB survey (Kalberla et al. 2005). For clarity the resolution was degraded to 2° . Contours are shown at column densities of 10 , 50 , and $120 \times 10^{18} \text{ cm}^{-2}$. Labels give the names of the cloud complexes



■ Fig. 12-3

Aitoff projection all-sky map of the column density of the intermediate-velocity gas with deviation velocity between $+35$ and $+90 \text{ km s}^{-1}$, based on the LAB survey (Kalberla et al. 2005). For clarity the resolution was degraded to 2° . Contours are shown at column densities of 5 , 10 , 50 , and $120 \times 10^{18} \text{ cm}^{-2}$. Labels give the names of the cloud complexes

2.3 Features of the HVC Sky

Wakker and van Woerden (1991) used the HVC surveys of Hulsbosch and Wakker (1988) and Bajaja et al. (1985) to construct a catalogue of individual clouds, containing 561 objects. Since these surveys did not have full sky coverage (having $1^\circ \times 1^\circ$ and $2^\circ \times 2^\circ$ grids, respectively), many southern clouds and some small northern clouds were not included.

A noticeable feature of the HVC and IVC sky is that the high-velocity gas appears to concentrate in large clouds and complexes of clouds. Historically, these have been given names that are either descriptive or consist of letters. A, B, and C were the first discovered HVCs (B is now considered part of A); M, H, and K were named after their discoverers (Mathewson, Hulsbosch, Kerr); R was one of a series of features in a paper on outer Galactic spiral arms; GCP (also known as GP or as the “Smith Cloud”), GCN (or GN), and AC were named after their location in the sky (near the Galactic Center and Anti-Center); L was named after a constellation (Libra); D and G are close to C and H in the alphabet and in the sky; WA/WB/WC/WD/WE were so named because the complexes were discovered by Wannier, Wrixon and Wilson (1972) and catalogued by Wakker and van Woerden (1991); the Magellanic Stream appears related to the Magellanic Clouds; the Intermediate-Velocity Arch, Low-Latitude Intermediate-Velocity Arch and Pegasus-Pisces Arch (IV, LLIV, PP) were named by Kuntz and Danly (1996) and Wakker (2001) for their curved appearance. Finally, “gp” is the intermediate-velocity gas near the HVC complex “GP”, with the names “g1”, “g2” and “g3” for the three main clouds. Although in the main these complexes are well-defined, near their edges the exact outlines are sometimes vague. Further, for many small clouds their relation to a complex can be ambiguous.

For each cloud the following useful quantities can be observed: (a) (l, b, v_{LSR}) , the location and average velocity of the gas; (b) Ω , the cloud area; (c) σ , the dispersion between the velocities in the various directions toward which the cloud is seen; (d) T_{peak} , the brightness temperature of the brightest spot in the cloud; and (e) $m(\text{HI})$, the mass of the cloud assuming a distance of 1 kpc. $m(\text{HI}) [M_\odot \text{kpc}^{-2}] = 0.236 S [\text{Jy km s}^{-1}]$, with S the 21-cm flux integrated over velocity (see [Chap. 11](#) by Dickey in this volume). The actual HI mass is calculated from the relation $M(\text{HI}) = m(\text{HI}) D^2 M_\odot$, where D is the distance to the cloud in kpc. The conversion between brightness temperature ($T_B(v)$) and flux ($S(v)$) is given by $T_B(v) = \frac{\lambda^2}{2k} \frac{S(v)}{\omega}$, with ω the solid angle of the main telescope beam. For a gaussian velocity profile the integrated flux is $S = 1.064 S_{\text{peak}} W$, where W is the full-width-at-half-maximum (FWHM), and the factor 1.064 is really $\sqrt{\pi/4 \ln 2}$. Thus, for the Dwingeloo telescope (used for the LAB survey and having $\omega \sim (4\pi/8 \ln 2) (0.6^\circ)^2$), $m(\text{single beam}) \sim 78 T_{B,\text{peak}} (W/20 \text{ km s}^{-1}) M_\odot \text{kpc}^{-2}$.

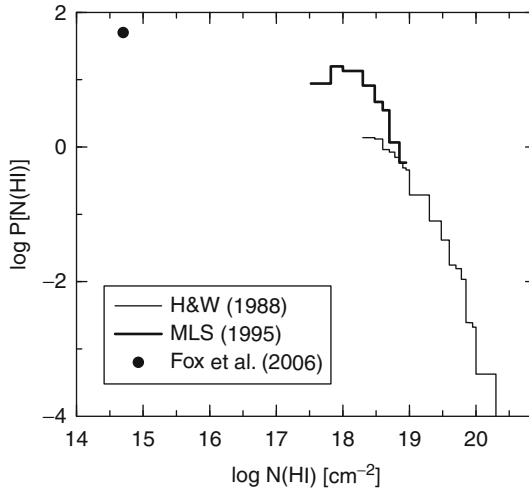
Wakker and van Woerden (1991) showed that although there are clearly recognizable HVC complexes, the distribution of the areas of individually defined clouds follows a power-law. That is, the number of clouds with area $>\Omega$ as function of area (Ω in square degrees) is

$$\log N(>\Omega) = -0.70 \log \Omega + 2.75$$

(using the updated values from Wakker 2004). Thus, for every 22 clouds with area >100 square degrees, there are 110 with area >10 square degrees, and 550 with area >1 square degree.

2.4 Sky Coverage

A question that is of substantial interest is toward how much of the sky high-velocity gas is seen. Giovanelli (1980) was the first to address this. Wakker (1991) studied it using the all-sky survey, and revisited it later (Wakker 2004). As [Figs. 12-1](#) and [12-4](#) show, the sky



■ Fig. 12-4


Percentage P of sky covered by HI with $v_{\text{DEV}} > 90 \text{ km s}^{-1}$. HI is binned in intervals of 10^{18} cm^{-2} for $N(\text{HI}) < 10^{19} \text{ cm}^{-2}$. For $N(\text{HI})$ between 10^{19} and 10^{20} cm^{-2} , the binning is in intervals of 10^{19} cm^{-2} , but the result is divided by 10. For $N(\text{HI})$ between 10^{20} and 10^{21} cm^{-2} , the binning is in intervals of 10^{20} cm^{-2} , but the result is divided by 100. This effectively makes all bins 10^{18} cm^{-2} wide, but smoothes out the sparsely populated bins at high column densities. The horizontal scale then displays $\log N(\text{HI})$ (see Wakker 2004 for a justification and more detailed explanation for using this kind of binning). The vertical scale gives the percentage of sky cover derived from three different datasets. The *thin solid histogram* is for data from Hulsbosch and Wakker (1988), the *thick solid histogram* for data from Murphy et al. (1995; MLS), while the large dot corresponds to the sky coverage seen in far-UV absorption by Fox et al. (2006)

coverage of gas with $|v_{\text{DEV}}| > 90 \text{ km s}^{-1}$ is substantial. In fact, HVC gas with $N(\text{HI}) > 10^{18} \text{ cm}^{-2}$ is seen toward 10% of the sky, most of it at negative velocities (7% vs 3% at positive velocities). Bright high-velocity gas ($N(\text{HI}) \gtrsim 10^{20} \text{ cm}^{-2}$) is rare, however, being seen in just $\sim 100 \text{ } 0.5$ beams (0.06% of the sky). On the other hand, fainter high-velocity gas ($N(\text{HI}) > 2 \times 10^{18} \text{ cm}^{-2}$) is seen toward 18% of the sky. No all-sky 21-cm surveys exist for even fainter clouds, but there are deep observations in both 21-cm and UV absorption in sightlines toward AGNs (QSOs and Seyfert galaxies). In particular, Murphy et al. (1995) observed 171 such directions. Using the detection fraction of high-velocity gas in that sample, Wakker (2004) estimated that the sky covering factor of HVCs with $N(\text{HI}) > 7 \times 10^{17} \text{ cm}^{-2}$ is 30%. The column density distributions found from the Hulsbosch and Wakker survey (1988) and implied by the Murphy et al. (1995) data are shown in ► Fig. 12-4.


Also included in that figure is an estimate of the sky coverage at much lower column densities, based on observations made using the Far-Ultraviolet Spectroscopic Explorer (FUSE). Sembach et al. (2003) detected high-velocity O VI absorption in 59 of 102 ($\sim 60\%$) sightlines to AGNs. O VI was seen in almost every direction where 21-cm HI emission was previously known, but also in many other sightlines. Fox et al. (2006) showed that 75% of the

O VI detections also show detectable H I in the Lyman absorption lines, with column densities down to $10^{14.7} \text{ cm}^{-2}$. They further find that the sky covering fraction of high-velocity H I with $N(\text{H I}) > 5 \times 10^{14} \text{ cm}^{-2}$ is on the order of 50%.

2.5 HVC Kinematics

A noticeable feature of  Fig. 12-1 is that at $0^\circ < l < 180^\circ$ most HVCs have negative deviation velocities, while at $180^\circ < l < 360^\circ$ most values of v_{DEV} are positive. This is even more pronounced when v_{LSR} is used (see Wakker 1991, Fig. 3a; Wakker and van Woerden 1991, Fig. 2b). This velocity asymmetry is caused by a combination of two effects. First, the LSR moves at 220 km s^{-1} toward $l = 90^\circ$, $b = 0^\circ$, so that clouds in that direction that are not participating in Galactic rotation tend to have negative velocities. Second, relative to the Milky Way as a whole, the maximum velocity of HVCs appears to be about 250 km s^{-1} (see below). Thus, a cloud near $l = 90^\circ$ that is receding from the Milky Way at $+250 \text{ km s}^{-1}$ will have an apparent observed velocity of $+30 \text{ km s}^{-1}$, and it will not be classified as an HVC. On the other hand, a cloud at $l = 90^\circ$ that is at rest relative to the Milky Way will appear as a HVC with $v = -220 \text{ km s}^{-1}$.

These two properties imply that the population of HVCs must be extended relative to the size of the Milky Way, and that on average the clouds are not taking part in Galactic rotation. If the clouds were local (distances less than a few kpc), their peculiar velocities would be very large relative to the denser material in the disk. If the total spread in velocities were larger than $\pm 300 \text{ km s}^{-1}$, negative velocity clouds would become visible near $l = 270^\circ$, and positive-velocity clouds would be seen near $l = 90^\circ$.

This effect is illustrated in  Fig. 12-5. The open circles in the left-hand panel show the longitude- v_{LSR} distribution of the clouds with $|v_{\text{DEV}}| > 80 \text{ km s}^{-1}$ in the Wakker and van Woerden (1991) catalogue. The solid points are the locations predicted from a simple model of the distribution of cloud locations and velocities, with red points for the 400 clouds that would observationally be classified as a HVC and blue points for the 447 objects in the population that are hidden by low-velocity emission. The panel on the left shows two observables and the panel on the right gives the locations and velocities projected onto the Galactic plane. The model has the following characteristics: (a) space density proportional to R^{-2} out to $R = 80 \text{ kpc}$ (with R the distance to the Galactic Center); (b) randomly oriented transverse velocity given by the Galactic potential at its location, such that the radial gravitational force is balanced by the centrifugal force; this velocity is on the order of $200\text{--}250 \text{ km s}^{-1}$; (c) a random transverse and radial velocity component that has a distribution with dispersion 50 km s^{-1} ; (d) a radial infall component of 50 km s^{-1} ; and (e) a similar population surrounds M 31 (see below).

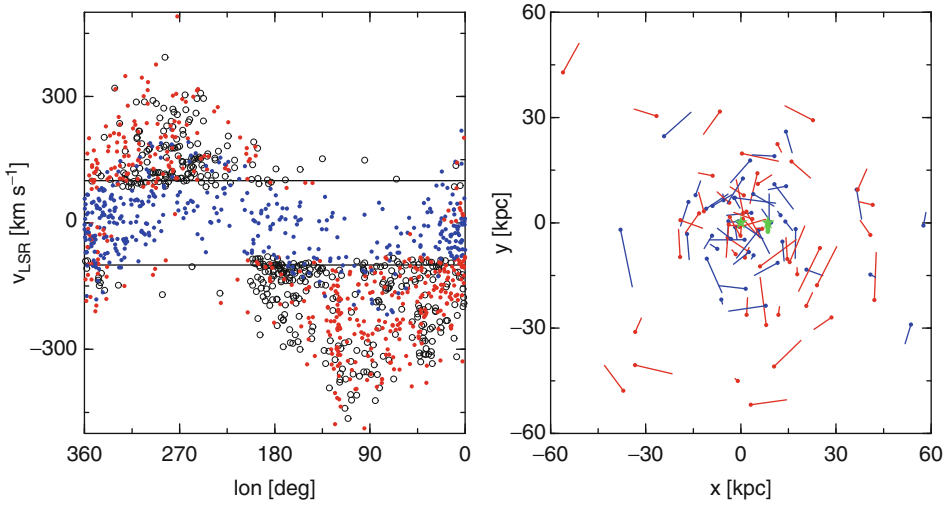
Defining a coordinate system (x, y, z) with the Galactic plane as the (x, y) plane and the Sun at $y = 0$, a cloud has a spatial location (x, y, z) and space velocity (v_x, v_y, v_z) , resulting in the following parameters:

$$r = \sqrt{x^2 + y^2}$$

$$d = \sqrt{(x - R_\odot)^2 + y^2 + z^2}$$

$$\Theta = \arctan \frac{y}{x}$$

$$l = \pi + \arctan \frac{y}{x - R_\odot}$$



■ Fig. 12-5

Left: comparison of observed and modeled distribution of the LSR velocity of HVCs as function of longitude. *Open circles* are for the clouds in the catalogue of Wakker and van Woerden (1991). *Red* and *blue* points are the predictions from the simple model described in the text, with red points for objects that would observationally be classified as a HVC. **Right panel:** locations and velocity vectors projected onto the Galactic plane for one fourth of the modeled clouds. The *green stars* give the locations of the Sun and of the Galactic Center

$$\begin{aligned}
 b &= \arcsin \frac{z}{d} \\
 v_r &= +v_x \cos \Theta + v_y \sin \Theta \\
 v_\Theta &= -v_x \sin \Theta + v_y \cos \Theta \\
 \sin \alpha &= \frac{R_\odot}{r} \sin l \\
 \cos \alpha &= \frac{r^2 + d^2 - R_\odot^2}{2rd} \\
 v_{\text{GSR}} &= (v_r \cos \alpha - v_\Theta \sin \alpha) \cos b + v_z \sin b \\
 v_{\text{LSR}} &= v_{\text{GSR}} + 220 \sin l \cos b,
 \end{aligned}$$

where r , d , Θ , and α are intermediate variables that go into the calculation of v_{GSR} (the cloud's velocity relative to the Galactic Center in a coordinate system in which the Milky Way rotates) and v_{LSR} (the cloud's velocity relative to the LSR).

The clouds can be given a mass based on a power-law mass spectrum:

$$N(M)dM = N_o M^\alpha dM; \quad M_{\text{low}} < M < M_{\text{upp}},$$

with $\alpha = -1.5$, as found by Wakker and van Woerden (1991). The upper mass limit is $10^7 M_\odot$, which is the mass found for HVC complex C (see ● Sect. 3). By assuming a constant density, n , the mass can be converted into a linear and angular size, which then allows deriving a brightness temperature. For the large, comparatively nearby HVCs typical densities of $0.05\text{--}0.15 \text{ cm}^{-3}$ are

found (see ● Sects. 2.6 and ● 4.3). For the more distant, less massive clouds that are embedded in a medium with lower pressure, a value of 0.01 cm^{-3} is more appropriate. Then:

$$\begin{aligned}
 R &= \left(\frac{3M}{4\pi m_a \bar{n}} \right)^{1/3} = 0.92 \left(\frac{M}{10^6 M_\odot} \right)^{1/3} \left(\frac{\bar{n}}{0.01 \text{ cm}^{-3}} \right)^{-1/3} \text{ kpc} \\
 \Omega &= \pi \left(\frac{R}{D} \right)^2 = \pi (318 \text{ arcmin})^2 \left(\frac{M}{10^6 M_\odot} \right)^{2/3} \left(\frac{\bar{n}}{0.01 \text{ cm}^{-3}} \right)^{-2/3} \left(\frac{D}{10 \text{ kpc}} \right)^{-2} \\
 S &= \frac{M(\text{HI})}{0.236 D^2} = \frac{2k}{10^{-26} \lambda^2} \sqrt{\frac{\pi}{4 \ln 2}} W T_{\text{B,p}} \Omega \Rightarrow \\
 T_{\text{B,p}} &= 0.59 \left(\frac{f}{0.5} \right) \left(\frac{M}{10^6 M_\odot} \right)^{1/3} \left(\frac{\bar{n}}{0.01 \text{ cm}^{-3}} \right)^{2/3} \left(\frac{W}{20 \text{ km s}^{-1}} \right)^{-1} \left(\frac{\min(\Omega, \Omega_{\text{beam}})}{\Omega_{\text{beam}}} \right) \text{ K}.
 \end{aligned}$$

Here M is the cloud mass in M_\odot , f is the neutral fraction ($f = M(\text{HI})/M(\text{H})$), m_a the average particle mass (~ 1.23 times the mass of a hydrogen atom, taking into account helium and heavy elements), R the cloud radius in kpc, D its distance to the Sun in kpc, Ω the area in steradians, S the flux integrated over the profile in Jy km s^{-1} , $\lambda \sim 0.21 \text{ m}$ the wavelength, W the FWHM of the velocity profile in km/s , and T_{B} the brightness temperature in K. The factor Ω_{beam} accounts for beam dilution. If the cloud is resolved, then T_{B} does not depend on D , as the apparent cloud area Ω also contains a factor D^{-2} . However, for unresolved clouds (i.e., clouds smaller than the telescope's beam size) beam dilution reduces the observed brightness temperature.

The simple kinematical model shows that the velocity distribution of the HVCs can be explained by a population of clouds that generally orbit the Milky Way, but that are also falling in with a net velocity of 50 km s^{-1} . This determines the velocity range from about -250 to $+200 \text{ km s}^{-1}$ when measuring velocities relative to the Milky Way, and -450 to $+300 \text{ km s}^{-1}$ when measuring relative to the LSR. About half of the clouds in such a population would not be classified as a HVC. The infall shifts the distribution in the l - v_{LSR} diagram such that the most negative velocity near $l = 90^\circ$ is 100 km s^{-1} more negative than the most positive velocity near $l = 270^\circ$. By themselves, these properties of the distribution cannot be used to determine the radial extent of the population, as long as it is much larger than the distance of the Sun from the Galactic Center. However, assuming that M 31 is also surrounded by a similar population of clouds implies an upper limit on the radial extent of about 80 kpc , since otherwise there would be an obvious concentration of HVCs in the sky area around M 31. Even in the model shown in ● Fig. 12-5, about 25 of the clouds are actually orbiting M 31. The detection of HVC-like objects near M 31 (Thilker et al. 2004) lends support to this picture of the HVCs.

2.6 Ionized HVCs

Not all of the hydrogen in HVCs and IVCs is in neutral form. In fact, it appears that H^+ represents a large fraction. This is to be expected, as can be shown by considering an atomic hydrogen layer of constant volume density n (in cm^{-3}) and fixed thickness (following Maloney 1993). First, define α , the hydrogen recombination coefficient ($\alpha = 2.6 \times 10^{-13} \text{ cm}^3 \text{ s}^{-1}$ for gas with $T \sim 10^4 \text{ K}$). Then, for a flux of ϕ ionizing photons $\text{cm}^{-2} \text{ s}^{-1}$ incident on both sides of this gas layer, there is a critical total hydrogen column density

$$N_c \sim \frac{2\phi}{\alpha n} = 7.7 \times 10^{18} \left(\frac{\phi}{10^{5.0}} \right) \left(\frac{n}{0.1} \right)^{-1} \text{ cm}^{-2},$$

below which the recombination rate ($\alpha n N_c$) is too small to balance the ionizing flux. Note that in the model of Bland-Hawthorn and Maloney (1999) $\log \phi = 4.5$ – 5.0 at galactocentric radii,

$R \sim 8\text{--}12$ kpc (for heights above the plane, z , up to 20 kpc); $\log \phi \sim 5.5$ at $R \sim 6$ kpc, $z < 10$ kpc; $\log \phi \sim 3.5$ at $R > 13$ kpc. Gas with column density $N < N_c$ will be mostly ionized, while gas with $N > N_c$ will be mostly neutral. The typical density of 0.1 cm^{-3} chosen above is based on the observed typical value in HVCs (see [Sect. 4.3](#)). The numerical value of the critical column density shows that most of the smaller, fainter HVCs probably are the central cores of larger, mostly ionized clouds. Mostly neutral gas is only expected to be present in the few percent of the sky covered by the brighter HVC cores.

In many circumstances, the H^+ can be observed by means of the $\text{H}\alpha$ photons that it emits. $\text{H}\alpha$ observations of HVCs are important for three reasons: First, they show the distribution of the ionized component of the clouds, which can be directly compared to the distribution of the neutral phase, revealing their full extent. Second, by combining $\text{H}\alpha$ and H I observations, it is possible to derive the volume density of a cloud. Adding data on the $[\text{S II}]$ optical emission line also allows the derivation of the temperature and pressure of a cloud. How to do this is described in [Sect. 4.3](#). Third, $\text{H}\alpha$ intensities contain information on the cloud distances and the intensity of the ionizing radiation surrounding the Milky Way; this is described in [Sect. 3.1](#).

Although $\text{H}\alpha$ emission has been detected in selected directions toward a number of individual clouds (Weiner and Williams 1996; Bland-Hawthorn et al. 1998; Tufté et al. 1998; Putman et al. 2003), maps are only available for two IVCs (complexes K and L – Haffner et al. 2001, 2005) and two HVCs (complex GP – Hill et al. 2009; complex A – Barger et al. 2012). An all-sky survey of HVC $\text{H}\alpha$ emission is not available, for two main reasons: First, to produce measurable $\text{H}\alpha$ emission requires gas with sufficient density, since the emission is proportional to the square of the density. Second, there are few instruments that have the sensitivity and velocity resolution to detect the HVCs, and the ones that exist have relatively narrow bandwidth, so that they need to be tuned to cover the range of velocities where high-velocity emission occurs.

3 The Determination of Cloud Distances and Metallicities

Two of the most important quantities that are used to understand the origin and properties of HVCs are their distance and their metallicity. Both of these are mainly derived from studies of absorption lines. This section describes the methods that are used to obtain information about HVC distances and metallicities.

By definition the HVCs do not take part in Galactic rotation. Therefore, it is not possible to combine their apparent velocity with a model of differential rotation to estimate a distance. Several different approaches have been proposed, which were summarized and critiqued in Wakker and van Woerden (1997) and in van Woerden and Wakker (2004). The most useful indirect techniques are first summarized, then the so-called absorption-line method is described in detail.

3.1 HVC Distances from Indirect Methods

3.1.1 $\text{H}\alpha$ Intensity

In principle the amount of $\text{H}\alpha$ emission coming from an HVC is a direct measure of the intensity of the ionizing radiation field, as on average each Lyman continuum photon is converted into $0.46\text{H}\alpha$ photons (see Spitzer 1998). Ferrara and Field (1994) work this out

in more detail, taking into account the confinement of HVCs by a surrounding medium, which affects the density structure, and thus the ionization structure and emission measure. Thus, in principle, if one has a prediction for the ionizing flux as function of location, then the observed H α emission measure directly gives the distance. Such a model was made by Bland-Hawthorn and Maloney (1999). However, when they applied their model to observations of the Magellanic Stream, they found that the observed H α emission was much stronger than was predicted, which means that an extra source of ionization is needed, or the model needs to be renormalized. Furthermore, for all HVCs that have been observed in multiple directions (complex K, L, GP, A, the Magellanic Stream), the observed H α intensity varies from point to point, which means that there are probably also geometrical (i.e., shadowing) effects due to small-scale structure in the clouds. It is more likely that in the end this method will be applied in reverse, to determine the intensity of the radiation field at different locations, using clouds with known distances.

3.1.2 Pressure Equilibrium

Since the clouds have a timescale for free expansion that is relatively short (see [Sect. 4.2](#)), they are either transient or confined by the pressure of a surrounding hot medium. Evidence for the latter comes from the detection of high-velocity O VI absorption, which is best interpreted as originating in an interface between the neutral cloud and a hot medium (see [Sect. 5](#)). Modeling the change of pressure with height above the Galactic plane can then give an estimate for the cloud's distance (Benjamin and Danly 1997; Espresate et al. 2002), giving values of a few kpc. However, the models are relatively simple and some of the necessary Galactic parameters are not well known.

3.1.3 Velocity Gradients

In a few larger clouds a velocity gradient is visible. Such gradients can be interpreted as due to projection effects, which then imply a distance. The most prominent case is that of the GCP complex, for which Lockman et al. (2008) find a distance of ~ 10 kpc, agreeing with the value derived from other methods. Using similar methods, Lockman (2003) derived a distance of 25 kpc for complex H.

3.1.4 The Virial Theorem

An easily derived quantity that suggests itself from the HVC observables (see [Sect. 2.3](#)) is the distance at which a cloud would be self-gravitating and in virial equilibrium. In that case

$$\sigma^2(3D) = \frac{2\alpha GM(\text{HI})}{fR},$$

where α is a factor near 1 that depends on the cloud geometry and virialization (Bertoldi and McKee 1992; see also Blitz et al. 1999); $\alpha = 1/2$ for a spherical cloud. f is the ratio of H I mass to total mass, including ionized and molecular hydrogen, helium and possible dark and stellar matter. For a cloud without dark or stellar matter in which hydrogen is 50% ionized,

$f = 0.5 * 0.748 = 0.374$, where 0.748 is the ratio $M(\text{H})/(M(\text{H})+M(\text{He})+M(\text{metals}))$. Converting the observables into the quantities in this formula ($\sigma(3D) = \sqrt{3} \sigma_{\text{obs}}$, $M(\text{HI}) = m(\text{HI}) D^2 = 0.236 S D^2 M_{\odot}$, $R = \sqrt{\Omega/\pi} D$) and rearranging yields:

$$D_{\text{vir}} = \frac{3}{2\pi^{1/2} \alpha} \frac{f \Omega^{1/2} \sigma_{\text{obs}}^2}{G m(\text{HI})} = 4500 \left(\frac{f}{0.374} \right) \left(\frac{\alpha}{0.5} \right)^{-1} \left(\frac{\Omega}{6sq^{\circ}} \right)^{1/2} \left(\frac{\sigma_{\text{obs}}}{10} \right)^2 \left(\frac{m(\text{HI})}{140} \right)^{-1} \text{ kpc}$$

A calculation of D_{vir} only makes sense for clouds that are detected in multiple beams. The typical mass of $140 M_{\odot}$ and typical area 6 square degrees are the median values for clouds with area >2 square degrees. Thus, the implied “virial distance” for a typical cloud is 4.5 Mpc.

The fact that such large distances are implied for clouds to be gravitationally stable if they only contain hydrogen used to be taken as an argument that they are unstable and transient (Hulsbosch 1975). Blitz et al. (1999), however, suggested that instead the HVCs are the visible part of the missing dark matter halos that are predicted to exist by cosmological simulations, with $f = 0.1$. They also assumed that a typical cloud structure gives $\alpha = 1$. This results in virial distances that are a factor ~ 10 smaller, on the order of a few 100 kpc, placing the HVCs in the Local Group. However, many more such halos are expected than the number of observed HVCs (see review by Kravtsov 2010).

Using this idea, Braun and Burton (1999) proposed that the small HVCs form a separate class, which they named the compact HVCs (CHVCs), which would be the Local Group objects, in contrast to the large, relatively nearby complexes. They mapped individual clouds in detail (see de Heij et al. 2002b and references therein), and in some cases interpreted the velocity field as due to rotation in a self-gravitating cloud. However, the area distribution of the HVCs follows a power law, suggesting that the CHVCs do not form a separate class. Further, not all CHVCs show a regular velocity field that can be interpreted as rotation.

In [Sect. 2.5](#) above, it was shown that the general population of HVCs is likely to have an extent of about 80 kpc, and that there are similar clouds associated with M 31. On the other hand, free-floating clouds with masses as large as those predicted in the Blitz et al. (1999) picture described above have not been detected in other galaxy groups that are similar to the Local Group, at least down to a detection limit of $4 \times 10^5 M_{\odot}$ (Zwaan 2001; Pisano et al. 2007). Maloney and Putman (2003) pointed out that, at distances on the order of 1 Mpc, the observed CHVCs would be so large that their HI densities would be on the order of $2 \times 10^{-4} \text{ cm}^{-3}$; hence, they would have to be predominantly ionized by the intergalactic radiation field, and their masses would be so great that the line widths would far exceed those observed. Thus, although many HVCs may be far from the Galactic plane, the population as a whole is unlikely to represent the missing dark matter halos at Local Group distances. This also implies that they likely are not self-gravitating, and thus some mechanism is required to hold them together.

3.2 Absorption-Line Method for Determining Distances

All methods listed above require major assumptions about the properties of the clouds and the Galactic environment, and they are not applicable in general. Statistically they might be useful, and they can be used as sanity checks. However, the only unambiguous results come from the “absorption-line method.” This method was described in several papers (Schwarz et al. 1995; van Woerden et al. 1999b; Wakker 2001). It requires the kinds of data described below.

3.2.1 Suitable Stellar Candidates Projected onto the HVC

A suitable type of star should be numerous at high galactic latitude, have a relatively reliable estimated distance, and have few stellar lines that can interfere with interstellar absorption. This leads one to Blue Horizontal Branch (BHB) stars, subdwarf B stars, and RR Lyraes. Before the year 2000 such stars were mostly found using low-resolution objective-prism spectra or photometric observations of individual stars (e.g., Kukarkin et al. 1970 – General Catalog of Variable Stars; Green et al. 1986 – the PG catalog; Beers et al. 1996 – the “HK” survey). At present, however, the great majority of candidate targets come from two surveys: 2MASS (Cutri et al. 2003), which provides JHK magnitudes for all objects in the sky down to magnitude $J \sim 15.5$, and the Sloan Digital Sky Survey (SDSS; York et al. 2000), which gives $ugriz$ magnitudes for QSOs and stars fainter than $g \sim 15$ in (mostly) the northern sky. Brown et al. (2004) determined the range of $J - H$ and $H - K$ colors of BHB stars from a sample of 550 spectroscopically observed halo stars, including 30 BHBs. They found that 65% of BHB stars have $-0.20 < (J - H)_o < 0.10$ and $-0.10 < (H - K)_o < 0.10$ and 41% of the stars in a sample selected using that criterion are BHBs. About 100,000 stars in the 2MASS survey fit these color criteria. In the case of the SDSS, there are several detailed studies that show how to use the $ugriz$ colors to select RR Lyrae and BHB candidates (Sirko et al. 2004; Ivezić et al. 2005). These studies yield a sample of 175,000 RR Lyrae and 15,000 BHB candidates, with about 85% accuracy. If a star was identified as a possible BHB or RR Lyrae, a preliminary distance can be estimated, using $M_V = 0.68$ for RR Lyraes (see [▶ Chap. 16](#) by Feast in this volume) and using the relation between absolute magnitude and color found by Preston et al. (1991) for BHBs. Next, these stars can be correlated with the high-velocity cloud catalogue to yield a list of stars at a range of distances in the direction of the HVCs.

3.2.2 Spectra and Photometry to Characterize the Candidates

It is possible to select the stars based on only their colors, but as the color-color selection is only about 75% accurate, it is necessary to obtain high-quality photometry and intermediate-resolution ($\sim 1 \text{ \AA}$) spectroscopy for a more detailed classification. The spectral shape, colors, and spectral features such as the width of the Balmer and Ca II H and K lines can then be matched against stellar atmosphere models to derive the stellar temperature, T_{eff} , gravity, $\log g$, and metallicity, Z . Comparing these numbers against theoretical isochrones (see e.g., Girardi et al. 2002) then yields an absolute magnitude. This also allows an estimate of the uncertainty in the absolute magnitude implied by the uncertainty in the stellar parameters.

3.2.3 High-resolution Spectra of Stars at Distances Bracketing that of the HVC

Absorption (by any ion) at the velocity of the HVC in the spectrum of a background star sets an upper limit on the cloud's distance, while a significant non-detection toward a foreground star sets a lower limit. A resolution of at least 15 km s^{-1} is needed, not only because that matches the typical width of the HVC absorption (thus optimizing the sensitivity), but also so that it is possible to separate the low- and high-velocity interstellar absorptions from each other and from stellar lines.

To convert a non-detection into a lower distance limit requires eliminating the possibility that it is due to too little HVC material in the direction of the star or to a low ionic abundance. This requires additional data: an accurate HI column density toward the star, and an accurate ionic abundance in the cloud. The latter can be obtained from a high-resolution spectrum of an extragalactic object or of a star known to lie behind the cloud, combined with good 21-cm data. The combination of a good ionic abundance and a good 21-cm column density allows a prediction of the equivalent width (EW) toward stars showing a non-detection. Then, as described by Wakker (2001), a lower distance limit follows if the ratio (predicted EW)/(observed 3σ EW limit) is larger than ~ 3 . For the UV lines of C II and Mg II the expected line strength is very large, with typical expected optical depth $\gg 1$, and spectra with low S/N ratio will suffice. However, this requires a telescope in space and the amount of available observing time is limited. For Ca II K (the best optical line; see [Sect. 3.4](#)), the typical expected equivalent width is 30–50 m Å, so a significant non-detection typically requires spectra with an equivalent width error of ~ 5 m Å, which requires an S/N ratio on the order of 50. Note that in the case of Ca II the typical expected equivalent width depends only very slightly on the HI column density (see [Sect. 3.4](#)).

To determine an upper distance limit to a cloud, it is sufficient to measure the equivalent width of an absorption line, which can be derived by a straight integration of the line profile:

$$W = \int_{\lambda_{\min}}^{\lambda_{\max}} \left(1 - \frac{F_{\lambda}}{C_{\lambda}}\right) d\lambda,$$

where F_{λ} is the observed flux, C_{λ} is the continuum flux, $d\lambda$ the wavelength step, and $\lambda_{\min/\max}$ are the wavelengths corresponding to the velocity range of the absorption.

To interpret non-detections requires an upper limit on the equivalent width, given by for example, three times the error. The total error in the equivalent width contains several contributions: photon counting noise, sky background, read-out error, dark current, the continuum fit, fixed-pattern noise, and uncertainties in the integration range. Spectroscopic calibration pipelines usually provide an error in the observed flux which combines the first four of these items. To convert this into an equivalent width error requires assuming a wavelength range over which to integrate, as the error is proportional to the square root of the integration range. The optimal choice is to match the width of the HVC 21-cm emission, typically 15–25 km s⁻¹. Combining in quadrature the flux errors with the errors in the continuum fit then gives a “statistical error.” The fixed-pattern error and the velocity-limits error can be combined in quadrature into a “systematic error.” In principle such a systematic error should also include uncertainties in the oscillator strength of the absorption line, but this is often ignored, and is usually relatively small. The statistical error indicates how accurately the measurement can be made, while the systematic error indicates how much the listed equivalent width could be offset from the actual value due to uncertainties associated with unknown but nonrandom offsets.

3.3 Considerations for Metallicity Measurements

In principle, measuring the metal content of a HVC is as simple as taking the ratio of the column density of a heavy element to the hydrogen column density. In practice there are the following items to take into account.

3.3.1 HI Small-Scale Structure

Heavy element column densities are usually measured by means of ionic absorption against background targets, most of which are QSOs or Seyfert galaxies (i.e., Active Galactic Nuclei or AGNs). Thus, there is a mismatch in angular sampling between the gas seen in absorption in the pencil-beam against the AGN and the gas seen in 21-cm emission using a radio telescope. This is minimized by taking radio spectra with the smallest possible beam, preferably using an interferometer such as the Westerbork telescope (WSRT), the Very Large Array (VLA), or the Australia Telescope (ATCA), which achieve beams of $<1'$. However, HVCs are often too faint to be detected by these interferometers. The Arecibo dish yields spectra with $3'$ resolution, but can only see part of the sky. The Green Bank, Effelsberg and Parkes telescopes have $\sim 10'$ – $15'$ beams, while the LAB survey covers the whole sky at $36'$ resolution. Wakker et al. (2001, 2011) find that using the LAB data yields column densities that are accurate to $\sim 25\%$, while using a $10'$ beam reduces the uncertainty to about 10%.

3.3.2 Depletion onto Dust

Dust is a ubiquitous constituent of the ISM, and its presence implies that the gas-phase column density measured for a heavy element does not represent the total column density. Savage and Sembach (1996) review this at length, comparing depletion in cold disk, warm disk, and halo gas, and they conclude that dust takes out fewer atoms in halo gas than in disk gas. However, except for oxygen and sulfur, depletion remains an issue. Where sulfur appears to avoid dust grains, oxygen will still be in the grains, but its large abundance means that most of the oxygen remains in the gas phase. In some circumstances silicon can also be useful, since in diffuse halo gas its abundance reduction is usually less than a factor 2. It should be noted that Savage and Sembach (1996) used measurements toward the IV-Arch and the PP-Arch to represent the halo gas. Note further that although dust is a problem when trying to estimate the metallicity of a cloud, estimates of the amount of dust in HVCs provide clues to their origin (see [Sect. 4.4](#)).

3.3.3 Presence of Ionized Hydrogen

As discussed in [Sect. 2.6](#), not all of the hydrogen in HVCs is neutral. To derive a metallicity therefore also requires an estimate of the relative content of ionized hydrogen, which is possible by measuring the cloud's $H\alpha$ emission, or by applying a photo-ionization model. Both of these methods require a distance estimate for the cloud, since the $H\alpha$ emission is proportional to the square of the density and the pathlength through the cloud, and the intensity of the ionizing radiation field depends on the cloud's location.

3.3.4 Elemental Ionization Fractions

In the diffuse ISM the balance between photoionization and recombination leads to a situation where most elements mainly occur in one dominant ionization state. Which state that is is

determined by the first ionization potential of the element that is larger than that of hydrogen, 13.6 eV. The relevant ionization potentials of sulfur, silicon, and iron are 10.36 eV ($S^0 \Rightarrow S^+$), 23.33 eV ($S^+ \Rightarrow S^{+2}$), 8.15 eV ($Si^0 \Rightarrow Si^+$), 16.35 eV ($Si^+ \Rightarrow Si^{+2}$), 7.87 eV ($Fe^0 \Rightarrow Fe^+$), 16.18 eV ($Fe^+ \Rightarrow Fe^{+2}$), which usually results in >90% of these three elements being in the form of S^+ , Si^+ , and Fe^+ in gas with total column density $>10^{18.5} \text{ cm}^{-2}$ (see next section). Since absorption lines from S^+ , S^{+2} , Si^+ , Si^{+2} , Fe^+ , and Fe^{+2} are all observable, this can often be checked observationally. Oxygen has an ionization potential that is close to that of hydrogen (13.62 eV for O^0 vs 13.60 for H^0), leading to a charge-exchange reaction that couples the column densities of O I and H I (Osterbrock 1989). This makes O I/H I the most reliable measure, as no correction is needed for the presence of O^0 , O^{+2} , or H^+ . That is, the fraction of ionized hydrogen and oxygen no longer matters, and the measured O I/H I ratio is only affected by dust depletion and small-scale structure (since the O I is measured in a pencil-beam, while H I is measured with a much larger 21-cm beam).

3.4 Strength of the Absorption Lines

The detectability of absorption lines from a given interstellar ion is determined by a combination of factors, including the element's cosmic abundance (A), the fraction of the element left in the gas phase after depletion onto interstellar dust grains (δ), the fraction of the element in a given ionization stage (F), and the oscillator strengths (f) and wavelengths (λ) of the absorption lines of the particular ion. Given a total hydrogen column density $N(H)$ (i.e., combining the neutral, ionized, and molecular phases) and an FWHM, W , the optical depth $\tau(\nu)$ as function of velocity associated with a given absorption line will be:

$$\tau(\nu) = \frac{\pi e^2}{m_e c} f \lambda_0 F \delta A N(H) \Phi(\nu)$$

where $\Phi(\nu)$ gives the absorption profile, normalized to have an integral of 1. That is, for a gaussian line profile the peak value of Φ , $\Phi(0)$ equals $1/(W \sqrt{\pi/4 \ln 2}) = 1/(1.064 W)$. The product $\frac{\pi e^2}{m_e c}$ has the value 2.654×10^{-2} in cgs units. Conventionally, however, λ is expressed in \AA , $N(H)$ in cm^{-2} , W in km s^{-1} , and Φ in $(\text{km s}^{-1})^{-1}$, in which case the coefficient has the value 2.654×10^{-15} .

• **Table 12-1** presents a simple estimate of the predicted optical depths and equivalent widths for a number of spectral lines given gas with solar abundance and total $N(H) = 10^{19} \text{ cm}^{-2}$ (assuming $W = 15 \text{ km s}^{-1}$). The lines listed include the most important lines in the far UV (O I, C II, Si II) and near UV (Mg II), which have been observed using the spectrographs on the Hubble Space Telescope. Also shown are results for the four strongest lines in the optical (the Ca II and Na I doublets).

The table shows that the UV lines of C II, Si II, and Mg II will typically saturate for H I column densities above $\sim 10^{18} \text{ cm}^{-2}$, making them easy to detect and in principle ideal to derive distance brackets for clouds. However, distant stars are usually too faint in the UV to observe with current space spectrographs. In the optical, the Ca II lines have an expected optical depth on the order of 0.1–0.5, making it possible to detect them in data with sufficiently high S/N ratio (>50). Note that Na I has not been detected in any HVC; considering the low expected optical depth this is not surprising.

Table 12-1

Relative strengths of absorption lines

Ion (1)	λ (Å) (2)	f (3)	$\log \delta$ (4)	$\log F$ (5)	$\log A$ (6)	τ_0 (7)	EW (m Å) (8)
O I	1302.169	0.0504	0.0	0.0	-3.31	53.4	165
O I	1039.230	0.00904	0.0	0.0	-3.31	7.65	98
C II	1334.708	0.1278	-0.30	0.0	-3.57	38.2	163
C II	1036.337	0.118	-0.30	0.0	-3.57	27.4	121
Mg II	2796.352	0.629	-0.38	0.0	-4.40	48.5	352
Mg II	2803.531	0.314	-0.38	0.0	-4.40	24.3	323
Si II	1260.442	1.070	-0.27	0.0	-4.49	38.9	158
S II	1250.584	0.00545	0.0	0.0	-4.88	0.149	9
S II	1253.811	0.01088	0.0	0.0	-4.88	0.299	18
S II	1259.519	0.01624	0.0	0.0	-4.88	0.448	26
Ca II	3934.777	0.6346	-1.0	-0.7	-5.66	0.181	36
Ca II	3969.591	0.3145	-1.0	-0.7	-5.66	0.091	19
Na I	5891.583	0.670	-1.0	-1.7	-5.76	0.023	7.1
Na I	5897.558	0.335	-1.0	-1.7	-5.76	0.011	3.6

Notes: Col. 1: Ion. Col. 2: Wavelength of line. Col. 3: Oscillator strength, from Morton (2003). Col. 4: Typical depletion for elements in the warm ISM; from Savage and Sembach (1996); for carbon, it is assumed that 50% is in dust. Col. 5: Typical ionization fraction for elements in the warm ISM at $N(\text{HI}) \sim 10^{19} \text{ cm}^{-2}$; these are implied by a calculation using the “CLOUDY” photoionization code (Ferland 1996); for O, C, Mg, Si, and S, F is generally ~ 1 , except in very low-density material ($n < 0.01 \text{ cm}^{-2}$), in high radiation fields ($\phi > 10^6 \text{ ph cm}^{-2} \text{ s}^{-1}$), and/or at low column densities ($N(\text{HI}) < 10^{19} \text{ cm}^{-2}$), where variations up to a factor 10 are possible; the value for Ca II is based on the empirical relation between the Ca II abundance and $N(\text{HI})$, using $N(\text{HI}) = 10^{19} \text{ cm}^{-2}$. Col. 6: Solar elemental abundance, from Asplund et al. (2009). Col. 7: Optical depth in the line center for an HI column density of 10^{19} cm^{-2} , assuming an intrinsic line width of 15 km s^{-1} and assuming solar abundances. Col. 8: Equivalent width of line for a total hydrogen column density of 10^{19} cm^{-2} .

Figure 12-6 presents a closer look at how the ionization of the important ions O I, S II and Ca II depends on the environment. This figure was created using the “CLOUDY” photoionization code (Ferland 1996). Column densities were calculated for a plane parallel layer, illuminated from one side, for a series of assumed volume densities ($n = 0.01\text{--}1 \text{ cm}^{-3}$) and several radiation fields (ϕ). In particular ϕ was chosen to represent the environment near the Sun (see caption). The spectral shape of the Milky Way radiation field was taken from Fox et al. (2004).

The figure shows that for a cloud located above the Galactic plane that has $N(\text{HI}) > 10^{18.0} \text{ cm}^{-2}$, all oxygen is in the form of O I, with deviations only occurring in intense radiation fields ($\log \phi > 6$), which could happen for an IVC with low HI column density that is close to the Galactic plane. For most HVCs the observed O I/H I ratios in HVCs lie below this line, showing that they have subsolar metallicity.

For sulfur, Figure 12-6 shows that ionization corrections may become important when $N(\text{HI}) < 10^{19.0} \text{ cm}^{-2}$, but only in low-density gas ($\log n < -1.0$) or in strong radiation fields ($\log(\phi/n) > 5$). Thus, the low S II/H I ratios that are observed in high column density HVCs indicate subsolar metallicity. In clouds with low $N(\text{HI})$ the situation is more complex. Fortunately, it is possible to use the S III $\lambda 1012.501$ line to assess this: if the physical conditions are

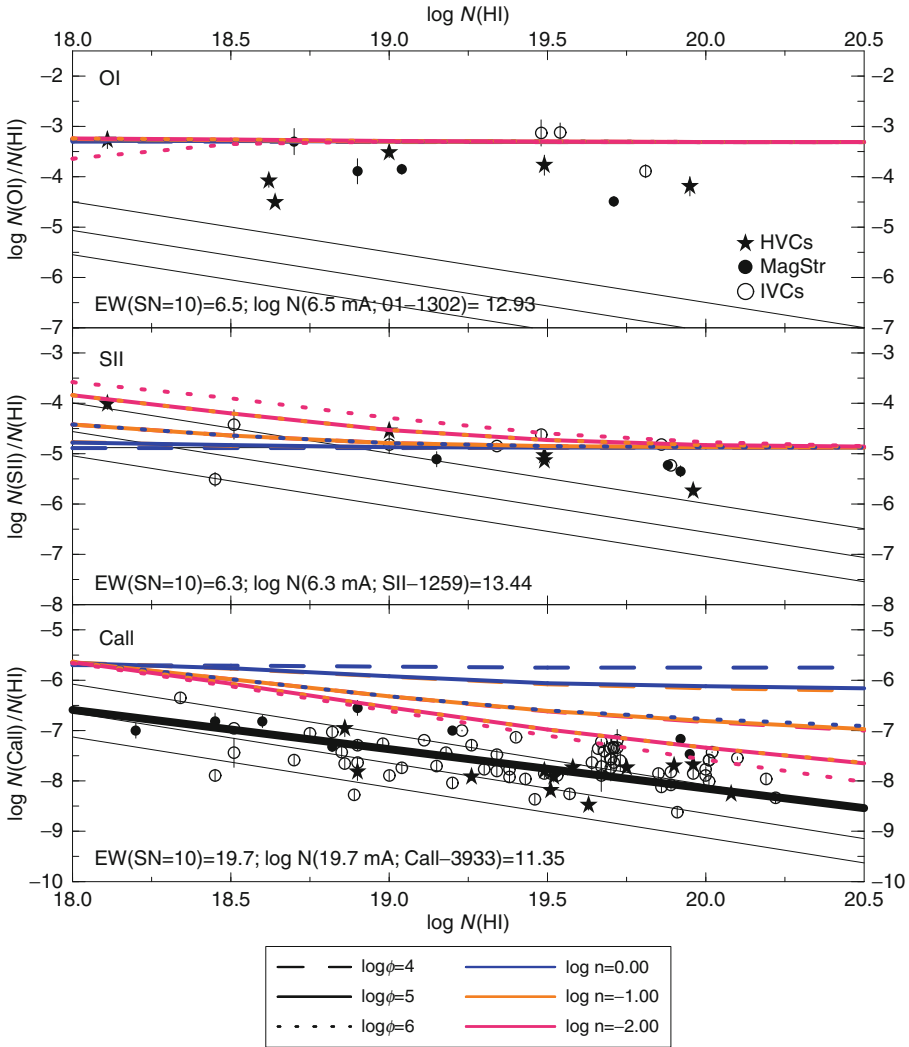



Fig. 12-6

Predicted apparent abundance ($N(\text{ion})/N(\text{HI})$) for O I, S II, and Ca II, as function of HI column density. Calculated using the “CLOUDY” photoionization code (Ferland 1996), for a plane parallel layer and assuming solar metallicity gas. *Solid lines* are for an ionizing radiation field intensity $\phi = 10^5 \text{ ph cm}^{-2} \text{ s}^{-1}$, corresponding to clouds at the solar circle with heights up to 20 kpc. *Dashed lines* are for $\phi = 10^4 \text{ ph cm}^{-2} \text{ s}^{-1}$, corresponding to locations outside the solar circle (from $R = 11 \text{ kpc}$, $z = 0 \text{ kpc}$ to $R = 18$ at $z = 10 \text{ kpc}$). *Dotted lines* are for a higher-intensity field in the plane and closer to the Galactic center (at $R \sim 3 \text{ kpc}$, $z < 3 \text{ kpc}$). Colors correspond to different volume densities. The *solid and dashed red lines* most likely represent the situation for HVCs. The observed values in HVCs, the Magellanic Stream and IVCs are given by filled stars, filled circles and open circles, respectively. Note that the observed O I and Ca II values lie below the colored model lines, which is due to the subsolar metallicity of the clouds in the case of O I, and due to depletion onto dust grains for Ca II. The three diagonal lines show the abundances corresponding to central optical depths of 1 (*top*), 0.1 (*middle*) and 0.03 (*bottom*), assuming a FWHM of 15 km s^{-1} . The *thick black line* in the Ca II panel is the empirical fit of Wakker and Mathis (2000)


such that $S\text{ II}/H\text{ I} > S/H$, then the S III line becomes detectable, and the S II/S III ratio can be used to determine the ionization corrections.

The behavior of Ca II differs from that of oxygen and sulfur because the ionization potentials of both Ca^0 and Ca^+ are less than 13.6 eV. Therefore, in gas with high H I column density, Ca^{+2} is the dominant ion, while Ca II/H I appears to be subsolar. In gas with low $N(\text{H I})$, Ca^{+2} is still the dominant ion at low volume densities, but Ca^+ is dominant at high volume density. The balance works out such that for $\log N(\text{H I}) < 18.3$ Ca II/H I is always about equal to Ca/H. For higher values of $N(\text{H I})$ a different situation can prevail, where in low density gas and/or high radiation intensity ($\log n < -1.5$ and $\log \phi > 4.5$) the Ca II column density varies by less than 0.4 dex for $\log N(\text{H I})$ between 18.0 and 20.0, so that the apparent abundance ($N(\text{Ca II})/N(\text{H I})$) decreases with increasing $N(\text{H I})$. As the datapoints in  Fig. 12-6 show, observationally the Ca II column density does indeed not vary much with $N(\text{H I})$ (as determined by Wakker and Mathis 2000). However, all the datapoints for Ca II lie below the colored model lines. In the case of the IVCs (open circles) the likely explanation is that calcium is heavily depleted onto dust grains, although this does not hide the trends in the Ca^+ ionization fraction. In the case of HVCs, the observed Ca II abundances lie below the model lines because the clouds have subsolar metallicity.

On average the following relation is found (see thick black line):

$$\begin{aligned}\log N(\text{Ca II}) &= 0.22 [\log N(\text{H I}) - 19.5] + 11.74 \\ \log A(\text{Ca II}) &= -0.78 [\log N(\text{H I}) - 19.5] - 7.76,\end{aligned}$$

(where $A(\text{Ca II})$ is defined as $N(\text{Ca II})/N(\text{H I})$).

This implies that on average the Ca II column density is $10^{11.8} \text{ cm}^{-2}$ for $N(\text{H I}) = 10^{20} \text{ cm}^{-2}$ and only decreases to $10^{11.4} \text{ cm}^{-2}$ when $N(\text{H I}) = 10^{18} \text{ cm}^{-2}$. Therefore, the Ca II H and K doublet provides a good way to determine cloud distances, as the lines will have similar strength in low and high column density gas. In addition, as the diagonal lines in  Fig. 12-6 show, the oscillator strength of the Ca II H and K doublet at 3934.777 and 3969.591 Å (vacuum wavelengths) is such that for gas with calcium abundance > 0.1 solar, the Ca II line is detectable in spectra with $S/N = 10$.

A final ion to discuss is Na I, as this has two strong absorption lines in the optical, at 5891.583 and 5897.558 Å. Na I is commonly detected in high column density low-velocity gas, and has been seen in a few IVCs, but not in HVCs. Its ionization balance is similar to that of Ca II, but the expected strengths of the absorption lines are such that in gas with low volume density the lines can only be detected in spectra with $S/N \sim 10$ if the gas-phase abundance of sodium is above ~ 0.25 solar, that is, the Na I doublet will generally not be detectable in clouds with subsolar metallicity and in clouds where a substantial fraction of the sodium is in dust grains.

3.5 Measured Distances and Metallicities

Wakker (2001) presented a comprehensive compilation of the published measurements of HVC absorption lines found in the spectra of stellar and extragalactic probes. That article also contains an extensive discussion of the distances and metallicities of individual HVC and IVC complexes, plus many maps and structural information. Since then, results from the *Far-Ultraviolet Spectroscopic Explorer (FUSE)* have provided much additional information concerning the metallicities of the HVCs. Many of those observations have been properly published (see Fox et al. 2004 and references therein), while others are only available in preliminary form

(as summarized by van Woerden and Wakker 2004). Further, *FUSE* detected H_2 in many IVCs, that is, in 50% of the directions where $\log N(\text{H I}; \text{IVC}) > 19.2$ (Wakker 2006). Finally, high-velocity O VI was found in >100 extragalactic sightlines (Sembach et al. 2003). To complement these UV results, observations with the Keck telescope, ESO's *Very Large Telescope (VLT)* and *Magellan* have yielded many additional distance brackets (Wakker et al. 2007; Wakker et al. 2008; Thom et al. 2006; Thom et al. 2008).

► **Table 12-2** summarizes available information on HVC distances and metallicities. Additional information for these clouds (such as mass and associated mass flows) can be found in Table 1 of Wakker (2004). ► **Figure 12-7** shows an example of stellar spectra that were used to derive a distance bracket to a cloud (core CIII in HVC complex C).

From ► **Table 12-2** it is clear that the distances to many of the larger HVC complexes (A, C, Anti-Center, GP) are on the order of 10 kpc, while the distances to the large IVCs (IV Arch, LLIV Arch, PP Arch) are only about 1 kpc. Some clouds appear to be more distant, however (clouds WW92, WW135, the Magellanic Stream). The cloud metallicities range from 0.09 times solar to about solar, with the values for complex C and the Magellanic Stream being the best determined. In some cases (complex A, complex WD), the derivation of a metallicity is complicated by the blending of O I $\lambda 1039.230 \text{ \AA}$ with lines of H_2 . It is clear, however, that the metallicity of the large IVCs is close to solar, while that of complex C is substantially subsolar, and that of the Magellanic Stream is similar to the value found in the Magellanic Clouds. Combined with the distances of ~ 1 kpc for the large IVCs, ~ 10 kpc for complex C, this is strong evidence for an explanation of the HVC phenomenon in which the large IVCs are related to the Galactic Fountain, complex C consists of low-metallicity accreting material, and the Magellanic Stream is a tidal stream pulled out of one or both of the Magellanic Clouds.

In the case of complex C an additional clue to its origin comes from the measurement of the deuterium to hydrogen ratio made by Sembach et al. (2004). They find $D/H = (2.2 \pm 0.7) \times 10^{-5}$. This value is (a) consistent with the primordial abundance of deuterium inferred from WMAP observations of the cosmic microwave background; (b) higher than that found for gas in the Galactic Disk, where deuterium is assumed to be destroyed inside stars, and (c) similar to values found in several QSO absorption line systems at redshifts >2 . Complex C is the *only* low-redshift cloud that has these three properties, and it thus provides an important anchor point for our understanding of the evolution of the D/H ratio.

4 Physical Properties of the HVCs

The internal structure of the HVCs gives important clues about their origin and fate. Relevant data include measurements of small-scale structure, velocity gradients, cloud size, temperature, density, pressure and timescales, ionization structure, and the relative amounts of neutral, ionized, and even molecular gas.

4.1 Small-Scale Structure

Maps of the HVCs have always shown structure down to the scale of the angular resolution. This is well illustrated by the case of complex A. With a $36'$ beam it appears to consist of several bright concentrations within a long filament (see ► **Fig. 12-1**). Observations with a $10'$ beam

■ Table 12-2

Known HVC distances and metallicities

HVC ^a	l, b_{cen}^b (deg)	v_{LSR}^c (km s ⁻¹)	Distance (kpc)	Dref	Metallicity (Z/Z _⊙)	Zref
Complex A (AIV)	155,38	-170:-140	>4.0	3		
Complex A (AVI)	160,45	-170:-140	2.0-10.0	9,29		
Complex C (CIA)	95,50	-150:-110	10.2-11.3	27	0.09	10,15,19,24,28
Complex C (CIIA)	130,54	-160:-140	7.4-10.9	23		
Complex C (CIIB)	115,54	-140:-110	<11.7	29	0.27	15,19
Complex C (CIIC)	120,58	-130:-110			0.15	14,19,20,24,28
Complex C (CD)	90,34	-190:-140	>12.6	27	0.19	18,24
Complex C (Cext)	65,38	-140:-120	<8.4	29		
Complex H	130,0	-220:-90	>5	6		
Mag. Stream		-430:+410	>30 ^d		0.25	7,11,12
ACII	195,-25	-150:-110	>2.7	25		
ACVHV (WW507)	168,-43	-350:-240	>0.3	16		
Cohen Stream (WW516)	145,-48	-170:-90	5.0-11.7	26		
Complex L	345,35	-190:-90	<4.5	29		
Complex GCP	40,-15	+90:+130	9.8-15.1	26		
Complex WA (WW135)	238,33	+100:+190	15-20	29		
Complex WB (WW92)	240,43	+90:+170	>20	29	0.2-0.6	21
Complex WD (WW226)	285,15	+110:+150			~0.1?	21
Complex WE	328,-15	+100:+120	<12.7	1		
Cloud WW35	249,52	+90:+120	6.4-8.3	22		
Complex M (MI)	165,65	-140:-110	3.4-4.3	29		
Complex M (MIII)	180,57	-120:-90	2.2-4.0	2,4,29		
Complex K	50,35	-100:-60	<6.8	16		
HVC100-7+110	100,-7	+100:+120	<1.3	8		
HVC224-83-197	224,-83	-190:-210			<0.5	17
IV Arch ^e	90,45	-95:-30	0.8-1.8	16	1.0	14
LLIV Arch ^e	150,32	-80:-30	0.9	16	1.0	13
Complex gp (g1)	67,-27	+55:+90	1.8-3.8	26		
Complex gp (g2)	35,-30	+55:+90	4.5	29		
PP Arch	120,-60	-80:-30	<0.9	16	~1.0	5

References: 1: Sembach et al. (1991); 2: Danly et al. (1993); 3: Wakker et al. (1996); 4: Ryans et al. (1997); 5: Fitzpatrick and Spitzer (1997); 6: Wakker et al. (1998); 7: Lu et al. (1998); 8: Stoppelenburg et al. (1998); 9: van Woerden et al. (1999a); 10: Wakker et al. (1999); 11: Gibson et al. (2000); 12: Sembach et al. (2001); 13: Richter et al. (2001a); 14: Richter et al. (2001b); 15: Gibson et al. (2001); 16: Wakker (2001); 17: Sembach et al. (2002); 18: Tripp et al. (2003); 19: Collins et al. (2003); 20: Sembach et al. (2004); 21: van Woerden and Wakker (2004); 22: Thom et al. (2006); 23: Wakker et al. (2007); 24: Collins et al. (2007); 25: Smoker et al. (2007); 26: Wakker et al. (2008); 27: Thom et al. (2008); 28: Shull et al. (2011); 29: Wakker et al. in preparation

^aCloud names WW # refer to the catalog of Wakker and van Woerden (1991)

^bApproximate Galactic longitude and latitude of center of complex

^cRange in LSR velocities

^dThe distance limit to the Magellanic Stream is based on numerical modeling

^eThe IV and LLIV Arch are the Intermediate-Velocity and Low-Latitude Intermediate-Velocity Arch, respectively

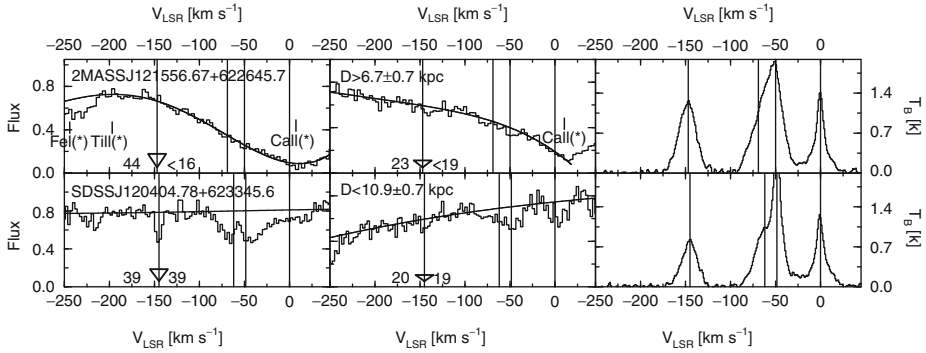


Fig. 12-7

Keck and LAB spectra for two stars that bracket the distance of core CIIIa in HVC complex C. Flux units are $10^{-16} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ \AA}^{-1}$. *Left column*: Ca II K spectra (histograms) and continuum fits (*solid curves*); *middle column*: Ca II H spectra and continua (note that the left flux scale is valid only for Ca II K); *right column*: HI-21 cm spectra from the LAB survey. One star per row, with the name of the star and the implied lower or upper distance limit given in the top left corner of the K and H panels. Labels (Ca II(*), Fe I(*), and Ti II(*)) show the positions of stellar absorption lines. *Vertical lines* give the velocities of the HI components. *Triangles* and numbers near the bottom axes in each Ca II panel give the expected HVC absorption line, the expected equivalent width, and the 3σ detection limit or the detected equivalent width. For the star on the top the expected Ca II K equivalent width is 44 m \AA , whereas the 5σ detection limit is 16 m \AA , leading to a significant non-detection and a lower limit of $6.7 \pm 0.7 \text{ kpc}$ to the distance of the HVC, where the error in the derived star distance is 0.7 kpc . For the star on the bottom, the HVC's Ca II K line is detected with equivalent width 39 m \AA , while Ca II H is seen at the 19 m \AA level. This sets an upper limit of $10.9 \pm 0.7 \text{ kpc}$ to the cloud's distance

(Giovanelli et al. 1973; Brüns et al. 2001; Lockman et al. 2008) reveal smaller cores within the concentrations. Synthesis telescopes such as the Westerbork Synthesis Radio Telescope (*WSRT*) and the Australia Telescope Compact Array (*ATCA*) provide maps with about $1'$ resolution, showing even more details (Schwarz and Oort 1981; Wakker and Schwarz 1991; Wakker et al. 2002; de Heij et al. 2002a; Schwarz and Wakker 2004).

A noteworthy feature of the structure seen at the highest resolutions is that in general the details show no apparent relation to the HVC as a whole, and no clear signs of interaction with the local gas. Further, the small-scale structure has random velocities within the velocity width of the profile at lower resolution. This may indicate that the small-scale features are short-lived condensations within the HVCs (see also Sect. 4.2 on timescales below).

Several approaches have been used to numerically characterize the small-scale structure. Some of these just aim at deriving a number that can be compared between different clouds. Others are inspired by theoretical models of the ISM, especially the idea that the structure is generated by turbulence. In general, none of these methods has yet been applied to a large number of clouds, because they all require having a large dynamic range in resolution, whereas most observations only cover about one decade.

A different approach is to use the autocorrelation function (or its Fourier transform, the power spectrum), which contains information on the average two-dimensional spatial size,

orientation, and amplitude distribution of all features within the map. Usually just the (one-dimensional) azimuthal average of the power spectrum is used, and this often results in a power law. Although it is not the only possibility, the most likely process for generating a power law is turbulence. Turbulence generates fluctuations in both density and velocity, in a particular manner, namely, by having energy injected into the medium at a large scale, forming large eddies that break up into ever smaller eddies until the injected energy is thermalized. Lazarian and Pogosyan (2000) showed how it is possible to use 21-cm data to disentangle the velocity and density fluctuations by analyzing the power spectra in velocity slices of varying thickness. However, to apply their methods requires data with a large dynamic range in size scales, so that so far this method only remains a promising way to analyze the structure of the HVCs when better data becomes available. A simple method that shows promise for understanding whether turbulence is a likely origin of the small-scale structure in HI clouds is to look at density statistics, as discussed by Burkhart et al. (2009).

4.2 Timescales

Significant understanding of the properties of the HVCs is provided by the several timescales that can be deduced from the measured sizes and velocities. These timescales were discussed in detail by Wakker and van Woerden (1991, 1997). They include the ones summarized below. All these timescale estimates are proportional to the cloud's distance, D . Below, a value of $D = 10$ kpc was used to obtain the typical values.

4.2.1 The Time it Will Take for a Cloud to Reach the Galactic Plane

This is found as

$$t_{\text{fall}} = \frac{z}{v_z} = \frac{D \sin b}{\sqrt{2} v_{\text{DEV}} \sin b} = 69 \left(\frac{D}{10 \text{ kpc}} \right) \left(\frac{v_{\text{DEV}}}{100 \text{ km s}^{-1}} \right)^{-1} \text{ Myr},$$

where z is the cloud's height above the plane and v_{DEV} its deviation velocity. The factor $\sqrt{2}$ is valid when assuming that the cloud's velocity perpendicular to the line of sight has a magnitude similar to its radial velocity (after taking out the projection of the motion of the Sun). The downward acceleration by the Milky Way's gravity can be neglected, which might reduce the time by about 10 Myr.

4.2.2 The Time for the Cores to Shift Substantially Relative to Each Other

$$t_{\text{shift}} = \frac{s}{\sigma} = \frac{\alpha D}{\sigma} = 42 \left(\frac{D}{10 \text{ kpc}} \right) \left(\frac{\alpha}{5^\circ} \right) \left(\frac{\sigma}{20 \text{ km s}^{-1}} \right)^{-1} \text{ Myr},$$

where s is the linear separation between cores, α the angular separation of the cores, and σ the internal velocity dispersion between the motions of different parts of the cloud. This relation only makes sense for the larger complexes with multiple cores. The value of 20 km s^{-1} is the median dispersion for about 60 clouds with surface area larger than 15 square degrees.

4.2.3 The Time a Core takes to Move Across its Own Width

This is the ratio of the core radius to the deviation velocity:

$$t_{\text{core}} = \frac{R}{v_{\text{DEV}}} = \frac{\alpha D}{v_{\text{DEV}}} = 1.7 \left(\frac{D}{10 \text{ kpc}} \right) \left(\frac{\alpha}{1^\circ} \right) \left(\frac{v_{\text{DEV}}}{100 \text{ km s}^{-1}} \right)^{-1} \text{ Myr},$$

where R is the linear radius of the core and α its angular size, typically 1° .

4.2.4 The Time for a Core to Double its Size

If the expansion were unrestrained:

$$t_{\text{expand}} = \frac{R}{\Gamma} = \frac{\alpha D}{\Gamma} = 8.5 \left(\frac{D}{10 \text{ kpc}} \right) \left(\frac{\alpha}{1^\circ} \right) \left(\frac{\Gamma}{20 \text{ km s}^{-1}} \right)^{-1} \text{ Myr},$$

where again R is the linear radius of a core, while Γ is the velocity dispersion inside a core, for which the width of the 21-cm emission line is a good approximation.

Even with the uncertainties in the distances and the rough estimates of sizes, the derived timescales for the processes that determine the small-scale structure are clearly much shorter than the lifetime of a whole complex. Thus, the relative location and internal structure of the cloud cores (timescales (c) and (d)) will change considerably during the movement of a complex through space (timescale (a)), and our present view is only a snapshot of a dynamic process. On the other hand, the cores will more or less stay in the same configuration as the cloud falls, since timescales (a) and (b) are similar.

4.3 Ionization Structure and Volume Density

Having a measurement of both H I and H α emission, as well as a distance, allows a derivation of the volume density and of the ionization fraction in a cloud, with the only remaining uncertainty being the assumed internal geometry. Define s as the coordinate along the line of sight, $n(s)$ as the volume density structure, and $x(s)$ as the ratio of ionized to total hydrogen. Further write the electron density as $n_e = \epsilon n(\text{H}^+) = \epsilon x n$. Unless the gas is hot enough to contain substantial amounts of ionized helium (ionization potential 24.6 eV, corresponding to $\sim 20,000$ K), $\epsilon \sim 1$. If helium is fully ionized, $\epsilon = 1.2$. The “standard” model has constant density, a fully neutral core and fully ionized envelope, that is, $x = 0$ in the core and $x = 1$ outside it. Other simple possibilities are to assume a gaussian density profile, and/or constant ionization throughout. The volume density can be rewritten as $n(s) = n_o n'(s/L)$, where n_o is the central density and L a length parameter giving the diameter of the cloud. Observationally, one can measure the angular diameter of the cloud (α). When assuming that the thickness of the cloud is the same as its width, then $L = \alpha D$, with D the cloud’s distance.

The intensity of the H α recombination emission coming from the ionized gas is measured in terms of Rayleigh (R), with $1 \text{ R} = 10^6/4\pi \text{ photons cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$ (see e.g., Haffner et al. 2003). But for a temperature-dependent factor, this is proportional to the emission measure (EM):

$$EM = \int n_e(s) n(\text{H}^+)(s) ds = 2.75 T_4^{0.924} I(\text{H}\alpha) \text{ cm}^{-6} \text{ pc},$$

where T_4 is the temperature T in units of 10^4 K. This temperature can be derived from observations of other optical emission lines, most notably [S II] $\lambda 6713$. In general, $T_4 \sim 1$.

With these definitions, the following relations hold:

$$\begin{aligned} N(\text{H}^+) &= \int x(s) n(s) ds = \mathcal{F}_1 n_o L \\ N(\text{HI}) &= \int (1 - x(s)) n(s) ds = (1 - \mathcal{F}_1) n_o L \\ EM &= \int \epsilon x^2(s) n^2(s) ds = \epsilon \mathcal{F}_2 n_o^2 L, \end{aligned}$$

where \mathcal{F}_1 and \mathcal{F}_2 are defined by these equations. Noting that $\epsilon \sim 1$, $T_4 \sim 1$, and $L = \alpha D$, and making a reasonable model for the density and ionization structure ($n(s)$ and $x(s)$) to give the structure factors \mathcal{F}_1 and \mathcal{F}_2 , these relations can be combined with the observables ($N(\text{HI})$, $I(\text{H}\alpha)$, α) to solve for the remaining two unknowns: n_o and $N(\text{H}^+)$. Observations of the forbidden [S II] emission line at 6713 \AA can be used to better constrain the temperature of the $\text{H}\alpha$ emitting gas (see Madsen et al. 2006 for a detailed description).

This method was applied to the H I, $\text{H}\alpha$ and S II absorption and emission data for HVC complex C in the sightline to Mrk 290 (Wakker et al. 1999). Inserting the recent measurement of the distance to this cloud (10 kpc) and assuming constant density and ionization fraction, this gives $n = 0.08 \pm 0.02 \text{ cm}^{-3}$, ionization fraction $x = N(\text{H}^+)/N(\text{H,tot}) = 17 \pm 10\%$, temperature $T = 7300 \pm 2000 \text{ K}$, and thermal pressure $P = 580 \pm 170 \text{ K cm}^{-3}$. Wakker et al. (2008) applied the same method to H I and $\text{H}\alpha$ for four clouds with known distance brackets, comparing the results using two different ionization models. In one model it is assumed that H^+ has the same pathlength as H I, in the other that $x = 0.5$ throughout. This resulted in volume densities in the range $0.05\text{--}0.15 \text{ cm}^{-3}$ and implied an ionized gas mass a factor 1–3 larger than the mass of neutral gas.

4.4 Molecules and Dust

Previous sections described the effects of dust on interpreting measurements of elemental abundances. The presence of and amount of dust in the clouds is also of intrinsic interest for two reasons: First, since dust usually forms due to stellar processes, the presence of dust in HVCs has implications for the history and origin of the gas. Second, through the well-established correlation between dust and molecules, dust gives information about the conditions in the cool cloud interiors. Direct searches for thermal dust emission from HVCs have been done using data from the Infra Red Astronomical Satellite (IRAS), with negative results for HVCs (Wakker and Boulanger 1986; Boulanger et al. 1996), but revealing emission from some IVCs (Désert et al. 1990; Weiss et al. 1999).

An indirect way to measure the presence of dust in HVCs is by comparing the abundances of heavy elements that are generally mostly in the gas phase (O, S) to those of elements that are generally present in the dust particles (e.g., Al, Fe, Ni). The analysis by Savage and Sembach (1996) shows that Si, Mg, Mn, Cr, Fe, and Ni are depleted by 0.3–0.8 dex in what they call “warm halo gas,” the measurements of which were done using several IVCs. Similar depletions were found in the LLIV Arch by Richter et al. (2001a). On the other hand, a detailed analysis of the relative abundances of different elements in complex C (Richter et al. 2001b) shows the absence of dust in this cloud, which suggests that it did not originate as gas in a stellar environment.

At the low densities typical for HVCs, molecular hydrogen (H_2) is only measurable by FUV absorption spectroscopy, requiring satellites in space, but then it can be seen at column densities as low as 10^{14} cm^{-2} . Using data from the FUSE satellite Wakker (2006) detects H_2 in 8 of 20 IVCs for which $\log N(\text{HI}) = 19.25\text{--}19.75$, and in 6 of 9 IVCs with $\log N(\text{HI}) > 19.75$, but in none of the IVCs for which $\log N(\text{HI}) < 19.25$. This is illustrated in [Fig. 12-8](#). Clearly, there is a transition from fully atomic gas to gas containing some H_2 at column densities above $2 \times 10^{19} \text{ cm}^{-2}$. On the other hand, the only H_2 at high velocity is seen in the Magellanic Stream, but not in 19 other sightlines, even though the median HI column density in these HVCs is $2 \times 10^{19} \text{ cm}^{-2}$.

No CO has been found in any HVC, even though deep searches were done toward selected dense cores (Wakker et al. 1997). In some IVC cores, however, CO has been detected. This includes the core IV21 in the IV Arch, toward which the intermediate-velocity HI is especially bright and narrow (Reach et al. 1994; Weiss et al. 1999), as well as several HI bright spots in the IV Spur (Magnani and Smith 2010). Since the CO emission is difficult to find and faint, no maps exist.

The hydrogen molecule is formed from neutral hydrogen when the volume density of the gas is sufficiently high (but only in the presence of a catalyst such as dust). H_2 is then destroyed by UV photons in the interstellar radiation field. Since this destruction takes away the photons, H_2 can survive in the denser central parts of a cloud. If there is an equilibrium between the formation and destruction of H_2 , then the following relation applies:

$$\frac{n(\text{HI})}{n(\text{H}_2)} = \frac{k\beta_0}{G T^{1/2} n(\text{H})},$$

where $n(\text{H}) = n(\text{HI}) + 2n(\text{H}_2)$ is the total volume density of protons. $k = 0.10 - 0.15$ is the probability that the molecule is dissociated after photon absorption, and β_0 is the photo-absorption rate per second. For a standard intensity of the interstellar radiation field

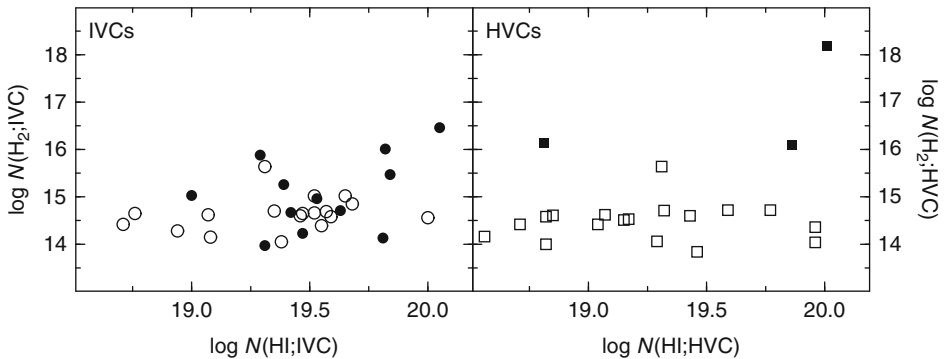


Fig. 12-8

Correlation between the column densities of atomic and molecular hydrogen in IVCs (*left*) and HVCs (*right*). Closed circles show detections, open circles are for upper limits. The detections of H_2 in HVCs are for a direction toward the Magellanic Stream (Fairall 9, Richter et al. 2001c), a direction toward the Leading Arm of the Stream (NGC 3783, Sembach et al. 2001), and a very small cloud at 75 km s^{-1} seen toward Mrk 153 (Wakker 2006). The 9 non-detections in HVCs with $\log N(\text{HI}) > 19.3$ include six sightlines through complex C, two through complex A, and one through the Outer Arm. The 13 detections in IVCs include 9 associated with the IV and LLIV Arch

($4\pi J_\lambda = 1.24 \times 10^{-5} \text{ W m}^{-2} \mu\text{m}^{-1}$; Mathis et al. 1983) $\beta_0 = 3.0 \times 10^{-10} \text{ s}^{-1}$. G is the probability per neutral H atom to form H_2 molecules by collisions with dust grains. In the Galactic Disk, $G = 9 \times 10^{-18} \text{ cm}^3 \text{ s}^{-1} \text{ K}^{-1/2}$ (van Dishoeck and Black 1986). This rate may differ in IVCs and especially in HVCs, however, as it depends on the presence of substantial amounts of dust, and because of the unknown surface properties of the dust grains in these clouds.

The problem with this equation is that it contains volume densities, while column densities are observed, and the H_2 will generally only be present in the regions with the highest densities. By defining χ as the strength of the interstellar UV radiation field relative to that near the Sun, and defining ψ as the fraction of the sightline where both H I and H_2 are present, the equilibrium relation can be turned into (Richter et al. 2003):

$$n_H \sim 1.2 \times 10^6 \frac{N(\text{H}_2)}{N(\text{H I})} \frac{\chi}{\psi} \text{ cm}^{-3}.$$

For the sightlines where intermediate-velocity H_2 is observed, the observed ratios of $N(\text{H}_2)/N(\text{H I})$ result in volume densities that range from $5/\psi$ to $35/\psi \text{ cm}^{-3}$, with a median of $27/\psi \text{ cm}^{-3}$. This implies pathlengths on the order of $\psi \text{ pc}$.

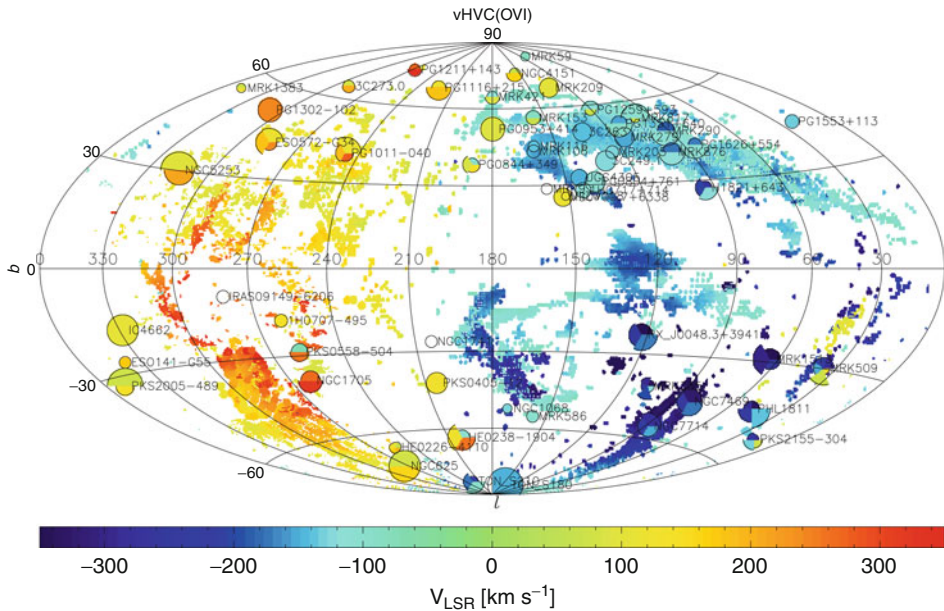
5 Hot Gas Associated with HVCs

In a 1956 paper, Spitzer pointed out that the lithium-like $(1s^2 2s)S_{1/2} \rightarrow (1s^2 2p)^2 P_{1/2,3/2}$ electronic transitions of O^{+5} , N^{+4} , C^{+3} , and Si^{+3} produce doublet absorption lines at wavelengths of 1031.9261/1037.6167 Å (O VI), 1238.821/1242.804 Å (N V), 1548.204/1550.781 Å (C IV), and 1393.7602/1402.7729 Å (Si IV). These ions have ionization potentials for production of 113.9, 77.5, 47.9, and 33.5 eV, which corresponds to temperatures in the range $1\text{--}3 \times 10^5 \text{ K}$. At these temperatures, interstellar gas is thermally unstable, as energy is carried off by photons emitted by collisionally excited ions, to which the medium is optically thin. At $T \sim 10^5 \text{ K}$ most of the energy loss is through resonance lines of oxygen (O IV, O V, O VI; see Sutherland and Dopita 1993). The presence of O VI, N V, C IV, and/or Si IV absorption therefore implies the existence of 10^6 K or hotter gas that has cooled down (or of a process that is heating up the gas).

The term “hot ISM” or “HIM” has often been used to describe gas with $T > 10^5 \text{ K}$, using “warm ISM” or “WIM” if the temperature is $\sim 5,000 \text{ K}$ to a few 10^4 K . However, since gas at temperatures $\sim 10^5 \text{ K}$ has very different properties than gas at $T > 5 \times 10^5 \text{ K}$ (not the least of which is that it cools relatively fast), Savage and Wakker (2009) proposed to instead use the term “transition temperature gas” for this quickly evolving phase, reserving “hot gas” for the X-ray emitting phase.

Starting in the 1970s, a sequence of spectroscopic facilities in space have observed the UV spectral lines of highly-ionized atoms. These include the Copernicus satellite, the International Ultraviolet Explorer (*IUE*), the Goddard High Resolution Spectrograph (*GHR*S), the Space Telescope Imaging Spectrograph (*STIS*), the Cosmic Origins Spectrograph (*COS*) (the latter three instruments on the Hubble Space Telescope), and the Far Ultraviolet Spectroscopic Explorer (*FUSE*). These instruments have found Si IV, C IV, N V, and O VI absorption in the Galactic Disk, with the most comprehensive studies those of Sembach and Savage (1992), Savage et al. (2001, 2003), Bowen et al. (2008), and Wakker et al. (2012).

Transition temperature gas at high velocity was first detected by Sembach et al. (1995, 1999) in the spectra of Mrk 509 and PKS 2155–304. The *FUSE* survey of O VI by Sembach et al. (2003) and the follow-up study of Fox et al. (2006) showed the ubiquity of this phase of the HVCs, as



■ Fig. 12-9

All-sky map showing the relation between the 21-cm HI high-velocity clouds (continuous background colors, cf. ● Fig. 12-1) and the high-velocity O VI absorption. A circle segment is shown for each sightline where high-velocity O VI is seen, with the color scale giving the velocities in the same manner as for the HI, and the radius of the segment proportional to the log of the O VI column density. Multiple circle segments in one direction indicate the presence of multiple absorption lines. Open circles are for directions without high-velocity O VI

high-velocity O VI was found in 85% of the 100 sightlines. ● Figure 12-9 compares the velocities of the O VI detections to the velocities of the 21-cm HVCs. With the possible exception of complex A, all 21-cm HVCs are found to have associated O VI absorption. High-velocity O VI is also seen away from the 21-cm clouds, with a large fraction of those cases occurring around the edges of large complexes such as complex C and the Magellanic Stream. Fox et al. (2006) found that in most of those cases (29 of 38, or 76%) there also is HI Lyman series absorption, with HI having column densities of 10^{15} – 10^{18} cm^{-2} . The average high-velocity O VI column density is $\log N(\text{O VI}) = 13.83$, and the distribution has a dispersion of 0.36 dex.

The presence of O VI in high-velocity clouds is most easily explained by concluding that they are embedded in a hot ($T > 5 \times 10^5$ K) surrounding medium. This is strongly supported by the fact that if both HI 21-cm emission and O VI absorption are seen, the velocities of these ions are well aligned (see Fox et al. 2004), but also by a comparison of predicted and observed column density ratios (see discussion below). That there is high-velocity O VI associated with the Magellanic Stream is therefore one of the strongest pieces of evidence for the existence of hot coronal gas around the Milky Way.

Several different mechanisms can produce highly-ionized gas in HVCs that are embedded in a hot corona. Numerical simulations give predictions for the column densities of the different

ions, as summarized below. Which processes are likely and unlikely can then be determined by comparing the predicted ionic ratios and column densities to the observed values.

For the following processes column density predictions have been published in the literature. These are compared with the observations in [Fig. 12-10](#). These models and their limitations are discussed in more detail by Wakker et al. (2012).

(1) *Collisional Ionization Equilibrium (CIE)*. In this case, at a fixed temperature every collisional ionization is balanced by a radiative recombination. Tables for the resulting ionic fractions were published by Sutherland and Dopita (1993) and by Gnat and Sternberg (2007). The predictions of these two papers are slightly different because of changes in atomic parameters (see references in Gnat and Sternberg 2007). Since many different processes affect the ISM (e.g., heating by supernovae, jostling by spiral arms and infall, cooling by photon emission), the gas is unlikely to be in CIE.

(2) *Shock ionization*. Shock fronts passing through the gas can create 10^5 K gas in which the highly-ionized atoms are present. Predictions for column densities were presented by Dopita and Sutherland (1996), for shock velocities of 200, 300, 400, and 500 km s⁻¹, magnetic parameter $B/n^{3/2} = 0, 1, 2$ and 4 $\mu\text{G cm}^{-3/2}$ (but only assuming solar abundances).

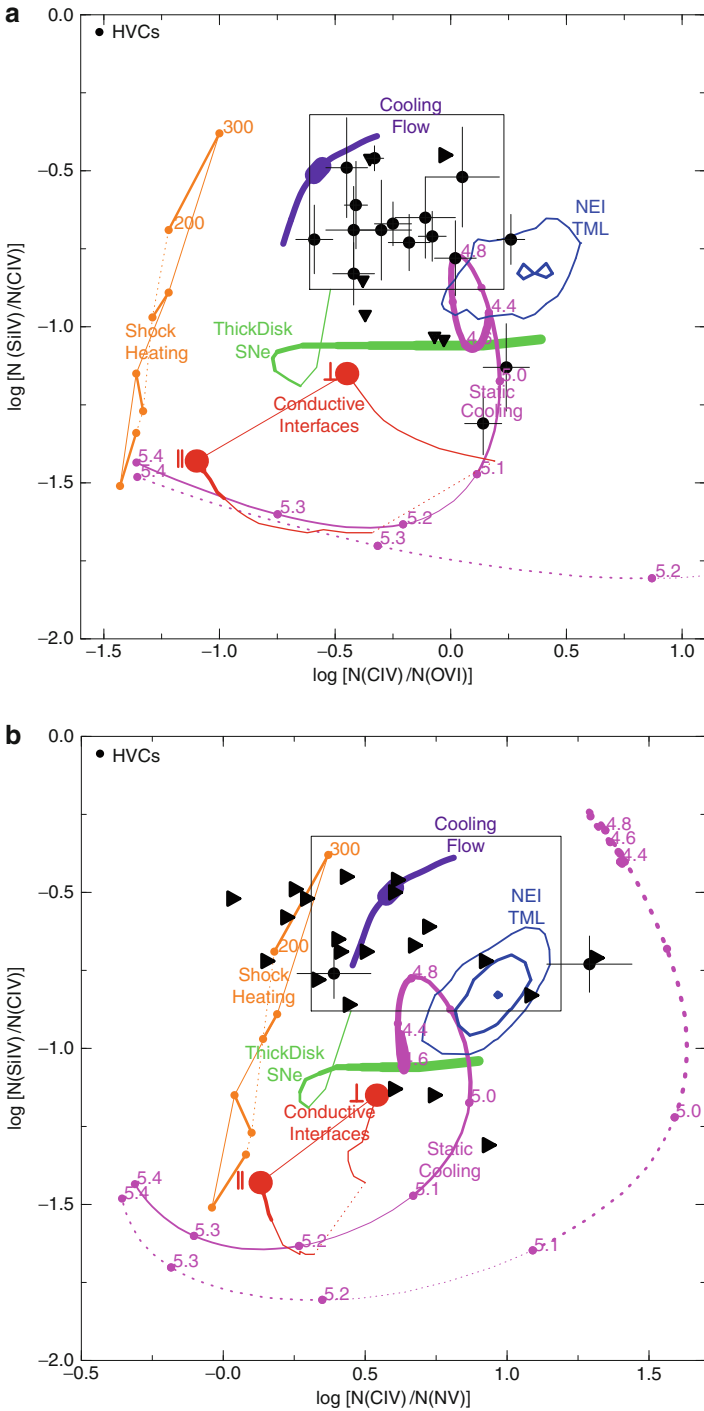
(3) *Static non-equilibrium radiative cooling*. In this kind of model a static parcel of gas starts out in CIE at a high temperature ($>5 \times 10^6$ K) and is allowed to cool, either isochorically (at constant density) or isobarically (at constant pressure). The resulting column densities were calculated by Gnat and Sternberg (2007). Most of the cooling is provided by metal-line emission, so the timescale is determined by the metallicity of the gas. At high metallicity the gas will cool faster than the highly-ionized atoms can recombine, leading to an overabundance of O VI, N V, and C IV relative to the amount present in CIE at the same gas temperature. Unlike what is the case for shock, conductive interface or turbulent mixing models, no absolute column densities are predicted by this model, only ratios.

(4) *Conductive interfaces*. If a region of cold (100 K) gas interfaces with a surrounding hot medium ($>5 \times 10^5$ K), electron thermal conduction transfers heat to the cold medium, creating gas at an intermediate temperature. After some time the heating will be balanced by radiative cooling, and a condensation front starts moving into the hot medium. Such interfaces were modeled in detail by Borkowski et al. (1990) and by Gnat et al. (2009). The models predict a thickness for the conductive interface on the order of 10 pc, with O VI and C IV column densities of about 5×10^{12} cm⁻².

(5) *Thick-Disk supernovae*. Shelton (1998) proposed that old supernovae that exploded in the Galactic Thick Disk (scaleheight 2 kpc) leave lingering amounts of highly-ionized gas that slowly recombines. She then predicted the time evolution of the column densities of the highly ionized atoms.

(6) *Turbulent Mixing Layers*. Gas at temperatures near 10^5 K can also be produced by turbulent mixing in shearing flows at the boundaries of hot and cold gas, as proposed by Begelman and Fabian (1990) and analyzed by Slavin et al. (1993), Esquivel et al. (2006), and Kwak and Shelton (2010). The latter of these presents a two-dimensional hydrodynamical calculation in which the non-equilibrium ionization of the different ions is followed. The earlier turbulent mixing models predict column densities in individual layers of $\sim 10^{12}$ cm⁻², much lower than what is observed, while the Kwak and Shelton (2010) models give $N \sim 10^{13}$ cm⁻². This implies that multiple interfaces are needed in a cloud to build up the observed median value of 7×10^{13} cm⁻².

(7) *Radiatively cooling gas flows*. If the cooling gas is moving, the thickness of the region where transition temperature gas is present will depend on the velocity of the flow. Detailed



■ Fig. 12-10
(continued)

models for this situation were summarized by Shapiro and Benjamin (1991), while more details can be found in the appendix included by Wakker et al. (2012). These authors also include photoionization due to photons generated in the hot phase, which is especially important for predicting the column density of Si IV.

Figure 12-10 compares the predictions to the observations for two pairs of ionic ratios, assuming that the gas has solar metallicity. In the panel on the top C IV/O VI is compared to Si IV/C IV, while the panel on the bottom gives the predictions for C IV/N V versus Si IV/C IV. Although these two ratios are similar, O VI is easier to detect than N V, and for many sight-lines high-velocity O VI has been measured. However, with existing missions only N V can be observed.

The predictions of the interface models (turbulent mixing, conductive interfaces, cooling flow) are expected to change if the gas has subsolar metallicity, as in that case cooling is less efficient. However, in models where the transition temperature gas is mostly formed by cooling (rather than by heating, as is the case for turbulent mixing), this should lead to a larger region where the cooling occurs. The increased pathlength then compensates for the lower efficiency, so that the resulting column densities of O VI, N V, and C IV should not be affected much. Unfortunately, except for the non-equilibrium cooling models of Gnat and Sternberg (2007) no detailed model calculations have been published for gas with subsolar metallicity, so this effect has not been directly confirmed.

Fig. 12-10

Theoretical ranges and observed values for three pairs of ionic ratios. DATA: Closed circles show ratios in individual HVC components; triangles show cases with one measured ratio and one upper limit. The square box encloses the range of observed ratios found in the Milky Way disk. MODELS: (1) Magenta lines (Static Cooling): non-equilibrium radiative cooling models from Gnat and Sternberg (2007), with $\log T$ labeled. The solid line is for solar metallicity gas, the dotted line for one tenth solar; the gas traverses this trajectory in about $3(0.001/n_0)$ Myr. In these models the gas is not flowing, and there is no additional heat input or output. (2) Orange lines (Shock Heating): Sutherland and Dopita (1993) shock models, with shock velocities of 200 and 300 km s^{-1} for the no-magnetic-field case marked and connected by a thick line. The other three thick line segments connect the 200 and 300 km s^{-1} case for higher magnetic parameters, $B n^{-3/2} = 1, 2,$ and $4 \mu\text{G cm}^{-3/2}$. (3) Red lines (Conductive Interfaces): conductive interface models from Borkowski et al. (1990), for perpendicular or parallel magnetic field (see the two symbols next to the red dots); the line thickness increases linearly with the age of the interface, 0.1 Myr at the thinnest point, 10 Myr at the other end, which is reached after 1–2 Myr. Thus, the line for this model appears as a quickly traversed thin red line ending in a large dot, after the ratios stabilize. (4) Green line (Thick Disk SNe): Shelton (1998) Thick Disk supernovae model, giving the evolution of the ratios with time (from 1 to 16 Myr after the supernova), with the line thickness proportional to the amount of time spent at each point. (5) Solid blue contours (NEI TML): distribution of ratios in the Kwak and Shelton (2010) turbulent mixing layer models, corresponding to 256 column densities derived by integrating through an interface at each 1 Myr long timestep between ages of 20 and 80 Myr. (6) Dark purple line (Cooling Flow): predictions of a model by Benjamin (see appendix in Wakker et al. 2012) for cooling hot gas flowing through an interface. The range of flow velocities is 14–16 km s^{-1} (thicker part of the line) and 9–26 km s^{-1} (thinner part)

Studies of the low-velocity absorption lines show that the absorption profiles of C IV and Si IV tend to be very similar, as are those of O VI and N V. Unfortunately, because of the lower cosmic abundance of nitrogen (as compared to oxygen or carbon), N V has so far been convincingly detected in only one HVC. Among the set O VI, N V, C IV, Si IV, the last two ions have the lowest ionization potential and photoionization may be important in their production. Except for the “Cooling Flow” model, this has not been taken into account in the predictions.

❖ *Figure 12-10* shows that *none* of the standard models of physical processes can reproduce *both* the observed C IV/O VI and the Si IV/C IV ratio in HVCs. The C IV/O VI ratio excludes shocks as a likely origin for the transition temperature gas. Conductive interfaces and halo supernovae seem incompatible with the Si IV/C IV ratio, although photoionization effects (not taken into account in the modeling) are expected to preferentially enhance Si IV, which may alleviate the discrepancy. Another problem with the conductive interface models is that they predict total column densities that are a factor 10–20 lower than what is observed. Static radiative cooling in gas with solar metallicity gives C IV/O VI and C IV/N V ratios that are similar to observations, but in low-metallicity gas straight cooling results in very different ratios than what is observed in HVCs. On the other hand, turbulent mixing and cooling flow models make predictions for Si IV/C IV, C IV/O VI/, and C IV/NV/ that are close to the observed range, although there are still some discrepancies. Further, in both of these types of models the total column density in a single layer is similar to the observed values (see discussion in Wakker et al. 2012).

With these caveats, it is clear that (a) the presence of transition temperature gas in HVCs implies that they are embedded in surrounding hot ($T > 5 \times 10^5$ K) gas, (b) the physical conditions in the interfaces are not fully understood, but they strongly indicate that non-equilibrium radiative cooling in a flow or mixing layer is likely to be important for creating the transition temperature gas seen in HVCs.

6 HVCs in and Around Other Galaxies

Since HVCs are observed from inside the disk of the Milky Way, only their radial velocities can be measured. Further, distances are known for only about 15 clouds. Thus, observations of HVCs in other galaxies may help our understanding. Various attempts have been made; for reviews see Oosterloo (2004), Sancisi et al. (2008), and Fraternali (2010).

Such observations have their own limitations, however. For distant galaxies the sensitivity and linear resolution are poor. For nearby galaxies, the nature of the data will depend on the galaxy’s inclination. Face-on galaxies allow direct measurements of the vertical component of cloud velocities, and they will show the horizontal shifts of clouds relative to features in the galactic disk. Edge-on galaxies directly provide the distance of clouds from the galactic plane, and show the horizontal component of the cloud velocities. For galaxies with intermediate inclinations, a more intricate analysis is required. Below, data on high-velocity gas for a few well-studied, nearby galaxies of various inclinations are discussed first. This is followed by a summary of the limited information available for a larger number of galaxies, and by some tentative conclusions.

6.1 A Face-on Galaxy: M 101

The large spiral galaxy Messier 101 is 5 Mpc distant and has low inclination, about 18° . Westerbork observations by van der Hulst and Sancisi (1988), with 0.6 kpc resolution, showed two

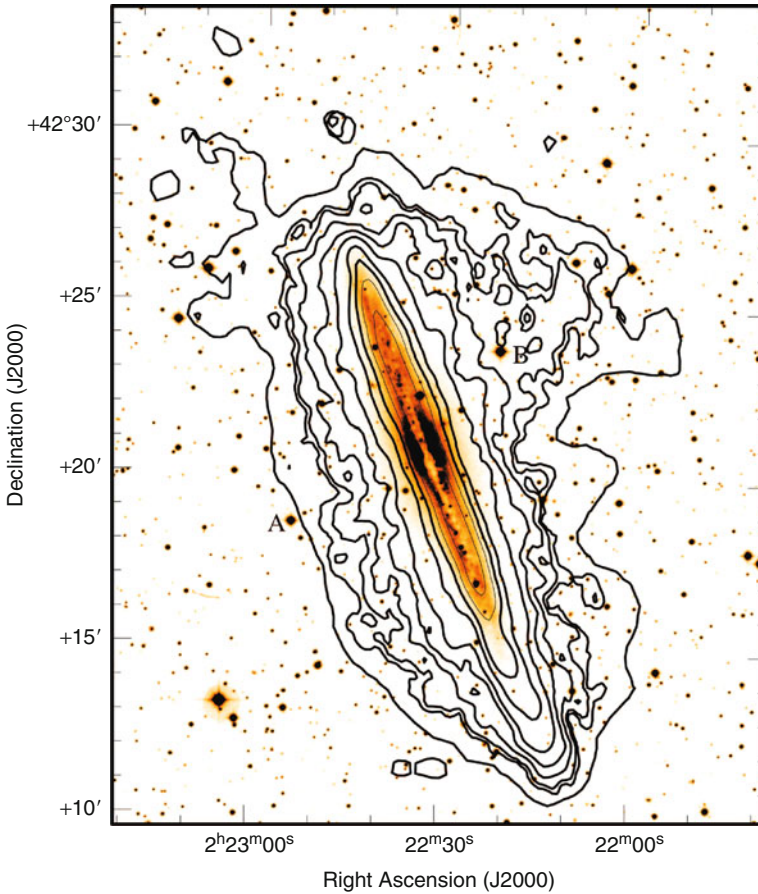
large HI clouds lying projected on M 101, with masses of 2×10^8 and $10^7 M_{\odot}$ and velocities of $+130$ and $+160 \text{ km s}^{-1}$ relative to the galaxy. Both clouds appear associated with a major hole in the galaxy's disk, evident also in position-velocity diagrams. The origin of the high-velocity clouds is not completely clear, but collisions of M 101 with large gas clouds are probable. A follow-up by Kamphuis et al. (1991) further describes an HI superbubble in M 101: a large (1.5 kpc diameter) hole in the disk, with gas moving away in two directions, at velocities of -50 and $+50 \text{ km s}^{-1}$. The shell has a mass of $3 \times 10^7 M_{\odot}$, and a kinetic energy exceeding 10^{53} erg, equivalent to 1,000 supernovae. This hole was first mentioned by Allen et al. (1973); it lies close to the H II complex NGC 5462, and may be related to it. The symmetry of motions suggests that the shell lies in the galactic plane, and is expanding into the halo.

6.2 An Edge-on Galaxy: NGC 891

This edge-on spiral galaxy lies at a distance of 9.5 Mpc, and has an inclination of at least 89° (Oosterloo et al. 2007). Early studies of H α spectra and images showed bubbles 0.5×1 kpc in size and diffuse H α emission up to $z = 2.4$ kpc (Pildis et al. 1994). Diffuse HI emission is found up to $z = 5$ kpc (Swaters et al. 1997). These studies also found that H α emission above the plane lags 40 km s^{-1} relative to the galaxy's rotation, while HI shows a lag of $25\text{--}100 \text{ km s}^{-1}$ (see also Heald et al. 2006a). Both the HI in the halo and the star formation in the disk were found to be stronger in the NE half of the galaxy. Deeper Westerbork observations by Oosterloo et al. (2007) (1.1×0.7 kpc resolution and HI detection limit $10^5 M_{\odot}$) show that the HI halo extends $10\text{--}15$ kpc from the plane, and locally even reaches $z = 22$ kpc (◆ Fig. 12-11). As much as 30% of the galaxy's HI lies in the halo. Oosterloo et al. (2007) also find halo clouds with mass $\sim 10^6 M_{\odot}$, some of which have "forbidden" (i.e., counter-rotating) velocities. Whether the halo HI is smoothly distributed or consists of clouds and complexes is not clear; this question requires higher resolution and sensitivity. Fitting 3-D models to Oosterloo's observations, Fraternali et al. (2005) find that purely ballistic fountain models fail, and that accretion of gas may be required. A 2-D rotation field shows a vertical rotation gradient of -15 km s^{-1} per kpc. Sancisi et al. (2008) note that these new observations have 50 times higher sensitivity than those of Sancisi and Allen (1979), but that the radial extent of the HI disk has not changed! Apparently, the HI halo extends only in the z -direction.

6.3 A High-Inclination Galaxy: NGC 2403

This nearby (3.2 Mpc), small spiral has an inclination of 63° . Following an initial analysis by Schaap et al. (2000), Fraternali et al. (2002a) analyzed VLA data with 0.2 kpc resolution (sensitive to clouds with $2 \times 10^4 M_{\odot}$) and found anomalous-velocity gas in position-velocity diagrams (the so-called *beard*; see ◆ Fig. 12-12). This is interpreted as originating from a 3 kpc thick HI component comprising $3 \times 10^8 M_{\odot}$ (10% of the galaxy's HI), that has a rotation lag of $25\text{--}50 \text{ km s}^{-1}$ in the inner parts, and an inward radial flow of 15 km s^{-1} . Long-slit spectra show that the ionized gas in the halo has the same rotation lag as the HI (Fraternali et al. 2004). Diffuse X-ray emission at 0.4–1 keV from the disk (described by Fraternali et al. 2002b) indicates a high star formation rate which probably is responsible for the disk/halo connection. All this fits with the results of Thilker et al. (1998) who identified 50 HI shells, determining their HI masses and kinetic energies. Combining all the evidence, the thick layer of HI and H α emission appears to be related to star formation in the disk through a Galactic Fountain.



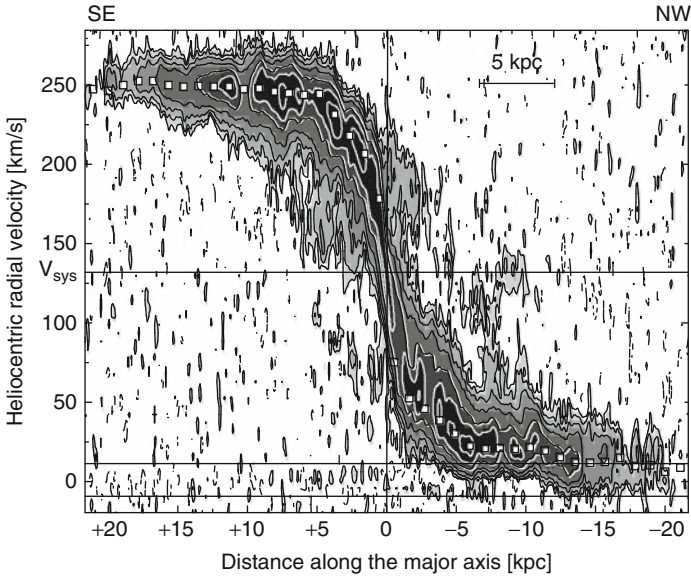
■ Fig. 12-11

Total H I image of NGC 891, obtained using the Westerbork Radio Telescope. The black contours are at 30'' resolution at levels of 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, and $50 \times 10^{20} \text{ cm}^{-2}$. The outermost contour is at a resolution of 60'' and a level of $0.05 \times 10^{20} \text{ cm}^{-2}$ (From Oosterloo et al. 2007)

In addition to this evidence for a Fountain, the position-velocity diagrams also show a large (8 kpc) high-velocity cloud, and gas with “forbidden” velocities up to 130 km s^{-1} , both appearing to be signatures of accreting gas.

6.4 A Low-Inclination Galaxy: NGC 6946

This large spiral, at 6 Mpc distance, has an inclination of 38° . It was studied by Kamphuis and Sancisi (1993) and Boomsma et al. (2008), with the latter study having 400 pc resolution. There is widespread high-velocity H I in this galaxy, moving at up to 100 km s^{-1} relative to the underlying disk. Much of the anomalous-velocity gas is seen near holes in the H I distribution, 121 of which were listed by Boomsma et al. (2008) (mostly in the inner disk). As in other



■ Fig. 12-12

Position-velocity diagram ($1'$ wide slice) along the major axis of NGC 2403 ($15''$ spatial and 10.3 km s^{-1} velocity resolution). The central horizontal line shows the systemic velocity, the other two horizontal lines mark the channels contaminated by HI emission from the Milky Way. Contours are $-0.26, 0.26, 0.5, 1, 2, 5, 10,$ and 20 mJy beam^{-1} . White squares mark the rotation curves for the two sides of the galaxy. The “beard” of anomalous-velocity gas refers to the emission that lies between the rotation curve and the systemic velocity at distances ~ -10 to $+10$ arcmin along the major axis. Forbidden velocity gas can be seen at $(d, v) \sim (+3, +75)$ and $(d, v) \sim (-5, +175)$ (From Fraternali et al. 2002a)

galaxies, the extraplanar HI lags in its rotation by $50\text{--}100 \text{ km s}^{-1}$. The anomalous-velocity gas represents a total energy of 10^{55} erg, most of which is probably contributed by stellar winds and supernovae. However, Boomsma et al. (2008) also invoked accretion to explain several peculiar phenomena in the outer disk, including velocity anomalies, sharp edges, and a strong asymmetry outside the optical disk. While the extraplanar HI amounts to $3 \times 10^8 M_{\odot}$, or 4% of the total HI in this galaxy, and shows clumpy and filamentary structure, its z -distribution remains unknown because of the galaxy’s low inclination (Sancisi et al. 2008).

6.5 The Nearest Spiral Galaxies: Messier 31 and 33

At 0.77 and 0.8 Mpc distance, the Andromeda Galaxy (M 31) and the Triangulum Galaxy (M 33) are by far the nearest large spirals, and therefore of great potential interest for the study of HVCs. However, M 31 (with an inclination of 75°) is strongly warped (Brinks and Burton 1984); hence, distances of objects from the disk plane are difficult to assess. Two HI surveys with a 100-m class telescope have been published, having 2 kpc linear resolution and a detection limit of $\sim 10^5 M_{\odot}$. Thilker et al. (2004) used the Green Bank Telescope to cover a 94×94 kpc region,

while Westmeier et al. (2008) covered the galaxy out to a radius of 100 kpc with the Effelsberg telescope. These maps reveal discrete features up to 50 kpc away from the disk, with HI masses of 10^5 – $10^6 M_\odot$, and velocities comparable to those of outer-disk rotation. There also is a filamentary halo component of at least 30 kpc extent concentrated at the systemic velocity of M 31. Using Westerbork, Westmeier et al. (2005) mapped 9 HVC fields at resolutions of 0.11–0.45 kpc, and identified 16 of the Thilker et al. (2004) clouds as HVCs, with masses of order $10^5 M_\odot$ and sizes ~ 1 kpc; 12 of these lie in a 1-degree area, within 15 kpc of the disk. The latter HVC complex partly overlaps with the giant stellar stream discovered by Ibata et al. (2001), and has similar velocities, suggesting a possible tidal origin. However, other elongated systems of HI clouds show no relationships with the additional stellar streams discovered around M 31 by McConnachie et al. (2009). Westmeier et al. point out that the dark-matter satellites predicted by CDM scenarios (Klypin et al. 1999; Moore et al. 1999) are not seen. This may be explained if the clouds are predominantly ionized. Finally, the region around M 33 was mapped by Putman et al. (2009) with Arecibo (linear resolution 0.8 kpc, sensitivity limit $\sim 2 \times 10^4 M_\odot$). They found five clouds similar to the Galactic HVCs, with HI masses between 3 and $20 \times 10^4 M_\odot$. Their origins might lie in tidal disruption of M 33 by M 31, a few Gyr ago.

6.6 High-Velocity and Extraplanar Gas in Other Galaxies

In addition to the well-studied galaxies described above, evidence for anomalous-velocity gas is found in many other galaxies (see e.g., Fraternali 2010 for a tabulation). In a number of galaxies the thick disk has been seen in both HI and H α emission. This includes NGC 891 and NGC 2403, and also NGC 5775 (see Rand 2000 and Heald et al. 2006a). Lee et al. (2001) further found extraplanar molecular and X-ray emitting gas in NGC 5775. In most if not all of these cases, the rotation of the extraplanar gas appears to lag behind that of the main disk. Generally, these galaxies are actively forming stars, suggesting an origin of the extraplanar gas as the result of supernovae. However, UGC 7321 (Matthews and Wood 2003) presents a special case that may suggest additional processes can be important. Although it is a low-surface brightness galaxy, it has a halo of 3 kpc thickness; the energy source for this extraplanar gas can hardly be sought in supernovae.

Additional evidence for a relation between anomalous-velocity gas and star formation comes from the study of HI holes and shells. Following the discovery of a hole near NGC 5462 in M 101 (Allen et al. 1973), later described as a supershell by Kamphuis et al. (1991), holes and shells have been found in many galaxies. Brinks and Bajaja (1986) discussed HI holes in M 31 of sizes 100–1,000 pc, representing HI deficits of 10^3 – $10^7 M_\odot$ and energies of 10^{49} – 10^{53} erg; ages were estimated at 2–30 Myr. Similarly, Deul and den Hartog (1990) found more than 100 holes in M 33. In both cases, smaller high-contrast HI holes appeared to correlate with OB associations and giant H II regions, while large holes have H II regions and OB associations on their edges. Two HI supershells in NGC 4631 are even larger (2–3 kpc, masses $\sim 10^8 M_\odot$ and energies exceeding 10^{55} erg; Rand and van der Hulst 1993a; Rand and van der Hulst 1993b). In the case of NGC 6946, Boomsma et al. (2008) directly associated the many HI holes with anomalous-velocity gas. The data on HI holes and (super)shells argue strongly that groups of early-type stars in the disk, and the associated supernovae, are a feasible energy source for galactic fountains and hence for extraplanar gas.

6.7 Accretion of Gas by Galaxies

As discussed in more detail in [Sect. 7.3](#), there may be two modes of accretion: “hot accretion,” in which clouds condense out of a hot virialized halo (dominant around massive structures), and “cold accretion,” in which streams of cold gas find a way down (dominant in star-forming galaxies with lower halo masses). Cold accretion may involve merging with gas-rich satellites or infall from the intergalactic medium. Sancisi et al. (2008) reviewed the accumulating evidence for “cold accretion” by galaxies, considering four types of observational evidence:

- (1) Gas-rich dwarfs, or HI complexes, tails and filaments, indicative of “minor mergers” of a galaxy with a smaller one, or of recent arrival of external gas.
- (2) Extraplanar HI found in nearby galaxies, undoubtedly in large part produced by galactic fountains, but probably also partly of extragalactic origin.
- (3) Warped outer HI layers.
- (4) Lopsided disk morphology and kinematics.

Types 3 and 4 represent more tentative evidence than types 2 and, especially, 1. Sancisi et al. (2008) tabulate 23 galaxies with dwarf companions and/or peculiar HI structures, and discuss and illustrate many of these in detail. The HI masses involved range from 1 to $50 \times 10^8 M_{\odot}$. The optical morphologies vary from undisturbed to asymmetric and peculiar. The HI has advantages as an accretion indicator: it provides a direct measure of the accreting gas, and its distribution and kinematics lend themselves well to merger modeling.

How frequent are these accretion events? Two studies suggest that ~25% of galaxies show signs of present or recent tidal interactions: an HI survey of a magnitude- and volume-limited sample of galaxies in the Ursa Major Cluster by Verheijen and Sancisi (2001), and the Westerbork Survey of HI in Spiral and Irregular Galaxies (WHISP, 300 galaxies) by van der Hulst et al. (2001). However, the lifetimes of the HI features in question are only of order 1 Gyr; hence, the frequency of past accretion events may be much higher than 25%. The HI masses and lifetimes mentioned suggest accretion rates between 0.1 and $1 M_{\odot} \text{ yr}^{-1}$.

7 Origins of the High-Velocity Clouds

In the first 30 years after their discovery in 1963 by Muller et al., many different theories were proposed to explain the anomalous-velocity Galactic gas (see review by Wakker and van Woerden 1997). Some of these proposed origins can now be excluded, while others are supported by the data, which show that: (a) Many galaxies have a thick layer of extraplanar HI and other gas; this includes our own Milky Way, where individual HVCs are seen several kpc above the plane. (b) The kinematics of the Galactic HVCs and the HVCs seen around M 31 indicate the existence of a population of clouds extending several tens of kpc from the disk. (c) There is evidence for accretion of gas from dwarf galaxies and/or intergalactic clouds, both from the distribution of HI seen around other galaxies, the existence of the Magellanic Stream and the low metallicity of several Galactic HVCs. (d) The Galactic HVCs are embedded in a hot surrounding medium. With these observations, it now is clear that the HVC phenomenon includes a mix of populations with multiple origins. Since the HVCs are observationally defined in very general terms, this multiplicity of origins is not surprising. However, three processes in particular are

thought to be responsible for the high-velocity and intermediate-velocity gas, and these will be discussed below.

7.1 The Galactic Fountain

High-mass stars put a large amount of energy into the ISM via radiation, stellar winds, and supernovae. Cumulatively, they create superbubbles and eventually chimneys, which vent hot gas into the halo, resulting in an extended density-stratified (at least on average) atmosphere around the Galaxy. Radial and vertical pressure differences then set up flows. This model was first put forward by Shapiro and Field (1976), who considered the implications of the presence of a Galactic corona that produces the observed soft X-rays and O VI absorption. Bregman (1980) then showed how conservation of angular momentum leads to an outward motion of the upwelling gas, and to an inward motion of the return flow. He proposed that this process can produce the HVCs. However, after further study (e.g., Mac Low et al. 1989; Houck and Bregman 1990; Rosen and Bregman 1995; de Avillez and Breitschwerdt 2005) it has become clear that the resulting apparent velocities are more similar to those seen for the IVCs. Bregman (2004) presents an extensive discussion of all the theoretical and observational aspects of the Galactic Fountain.

The temperature of the Galactic coronal gas is estimated to be $\sim 5 \times 10^5$ K by combining inferences from X-ray emission (see McCammon and Sanders 1990 and Snowden et al. 1998), the presence of transition temperature gas with a scaleheight of a few kpc (Savage and Wakker 2009), and CIV and O VI emission lines (Martin and Bowyer 1990; Shelton et al. 2007). The density of the 5×10^5 K gas in the disk, n_0 , can be estimated in the manner presented by Benjamin (2004), who shows the densities for an isothermal 10^6 K halo using the galactic potential model of Dehnen and Binney (1998). A more direct observational estimate comes from comparing X-ray and O VIII line absorption and emission (Bregman and Lloyd-Davies 2007). Both of these methods suggest that $n_0 \sim 10^{-3}$ near the Sun.

Assuming that there is a source of hot gas, the argument of Kahn (1976, 1981) reveals the physical conditions under which a fountain flow will develop. First define the mean atomic mass number $m_a = \sum_{\text{elements}} A_{\text{el}} m_{\text{el}} \sim 1.23 m_H$, with A_{el} and m_{el} the abundance by number and the atomic mass of each element, and m_H the mass of a hydrogen atom (1.67×10^{-27} kg; $A(\text{H}) = 0.925$ and $A(\text{He}) = 0.074$). Second, define the mean particle mass, m_p , which includes counting the electrons. In fully ionized gas ($T > 10^5$ K) $m_p \sim 0.5 m_a$. m_a is used to calculate the mass density $\rho = m_a n_H$, while m_p is used to calculate the pressure, $P = \frac{\rho}{m_p} kT$. Finally, near the Sun, the gravity above the stellar layer is $g_z = 6.25 \times 10^{-9}$ cm s $^{-2}$ (Kuijken and Gilmore 1991).

Given these definitions, the sound speed c_s is given by:

$$c_s = \left(\frac{\gamma P}{\rho} \right)^{1/2} = \left(\frac{\gamma k}{m_p} \right)^{1/2} T^{1/2} \sim 67 \left(\frac{T}{10^{5.3}} \right)^{1/2} \text{ km s}^{-1}$$

where $\gamma = 5/3$ is the ratio of specific heats. Thus, the natural scaleheight of the hot ISM is $H = c_s^2/g_z \sim 2.3(T/10^{5.3})$ kpc. The dynamical timescale for the gas flow is given by:

$$t_{\text{dyn}} = \frac{c_s}{g_z} \sim 34 \left(\frac{T}{10^{5.3}} \right)^{1/2} \left(\frac{6.25 \times 10^{-9}}{g_z} \right)^{-1} \text{ Myr.}$$

Kahn (1976) showed how to estimate the cooling timescale, by calculating the rate of energy loss per unit mass (Q) and comparing this to the rate of change of specific entropy (s): $Ds/Dt = Q/T$. This is done by approximating the standard interstellar cooling function (Λ in units of $\text{erg cm}^3 \text{s}^{-1}$) at temperatures above 10^5 K by a power law. Then:

$$Q = \frac{\Lambda n_H}{m_a} \text{ erg s}^{-1} \text{ g}^{-1} \text{ with}$$

$$\Lambda = 1.33 \times 10^{-19} T^{-1/2} \text{ erg cm}^3 \text{ s}^{-1}.$$

Given a volume V and a number of particles N , the entropy of a monatomic ideal gas is:

$$S = Nk \log \left[\frac{V}{N} \left(\frac{U}{N} \right)^{3/2} \right] + \text{Constant}.$$

Inserting $U = (3/2)NkT$ for the internal energy, using $M = Nm_p$, and defining $\kappa = P\rho^{-5/3}$, the specific entropy ($s = S/M$) can then be written as:

$$s = \frac{k}{m_p} \log \left(\kappa^{3/2} \right) + \text{Constant}.$$

Using κ , and using that $Ds/Dt = Q/T$, it can be shown that:

$$\frac{D\kappa^{3/2}}{Dt} = \frac{-1.33 \times 10^{-19} k^{1/2}}{m_a^2 m_p^{1/2}} = -q,$$

with $q \sim 3.6 \times 10^{32} \text{ cm}^6 \text{ g}^{-1} \text{ s}^{-4}$. Since q is a constant, this gives the cooling time due to radiative losses as:

$$t_{\text{cool}} = \frac{\kappa^{3/2}}{q} = \frac{P^{3/2}}{q\rho^{5/2}} = \frac{k^{3/2}}{q m_p^{3/2} m_a} \frac{T^{3/2}}{n_H} = 5.8 \left(\frac{T}{10^{5.3}} \right)^{3/2} \left(\frac{10^{-3}}{n_H} \right) \text{ Myr}.$$

The ratio of the cooling timescale to the dynamical timescale is therefore:

$$\frac{t_{\text{cool}}}{t_{\text{dyn}}} = \gamma^{-1/2} \frac{P}{\rho^2} \frac{g_z}{q} \sim 0.17 \left(\frac{T}{10^{5.3}} \right) \left(\frac{10^{-3}}{n} \right).$$

Thus, near the Sun, the cooling time is much shorter than the dynamical time. Therefore, if there is a source of hot gas, it will cool before it reaches hydrostatic pressure equilibrium (with a scaleheight of about 2–3 kpc). The cool gas will lose buoyancy and rain down to the plane. Since the gravity, density, and temperature of the hot gas vary considerably across the Galaxy, a Fountain flow may not be present everywhere, however. Near the Galactic Center, the situation changes since the gravitational potential is no longer plane-parallel and the gas temperature is higher. Then a Galactic Wind may occur (e.g., Breitschwerdt et al. 1993; Everett et al. 2008).

The identification of at least some IVCs (specifically the IV Arch and LLIV Arch) with the return flow of a Fountain is based on their near-solar metallicity (showing they originated in the disk), the fact that their dust content is lower than that of gas in the disk (suggesting the dust was partially destroyed when the gas was ejected from the disk), their height above the plane ($\sim 0\text{--}2 \text{ kpc}$) and their velocities (about 50 km s^{-1} downward motion). There is also evidence for mostly ionized gas at $z \sim 1 \text{ kpc}$ that is flowing up (cloud g1, Wakker et al. 2008). However, for many other IVCs the metallicity and distance have not yet been measured, so it remains unclear whether these would also represent Fountain clouds.

To derive the mass flow rate corresponding to these clouds requires an estimate of their vertical velocity, which is not directly observed, but which can be estimated. A reasonable first guess can be made by using v_{DEV} , assuming that the tangential component of the velocity vector is equal to the radial component, and projecting v_{DEV} . Then, summing the column density ($N(\text{HI})$) over each telescope beam (with size $d\Omega$) that covers the cloud:

$$\dot{M} = \frac{Mv_z}{z} = \int \sqrt{2} v_{\text{DEV}} \sin b m_{\text{H}} N(\text{HI}) D d\Omega,$$

where m_{H} is the mass of an H atom and D the cloud distance. Using the parameters for the IV and LLIV Arches ($v_{\text{DEV}} = 50 \text{ km s}^{-1}$, $b = 45^\circ$, $N(\text{HI}) = 10^{20} \text{ cm}^{-2}$, $\Omega = 5,000$ square degrees, $D = 1 \text{ kpc}$) shows they represent a circulation rate of $\sim 0.06 M_{\odot} \text{ yr}^{-1}$ within 1 kpc of the Sun. Assuming that the Fountain covers $\sim 25\%$ of the Galaxy (300 kpc^2 out to $R = 10 \text{ kpc}$), the total circulation rate would be about $4.5 M_{\odot} \text{ yr}^{-1}$.

A completely independent way of estimating the mass flow rate comes from the intensity of the C IV emission line, which is proportional to the product of the density of the ionized gas (n , in cm^{-3}) and the velocity of the flow (v , in km s^{-1}). Observations (Martin and Bowyer 1990) imply $n v = 1.8 \times 10^5 \text{ cm}^{-2} \text{ s}^{-1}$. Theoretically, the circulation rate is simply given by the equation of mass conservation. Translating to quantities useful for Galactic studies, the two-sided mass flow rate is

$$\dot{M} = 4.3 \left(\frac{f}{0.25} \right) \left(\frac{nv}{1.8 \times 10^5 \text{ cm}^{-2} \text{ s}^{-1}} \right) \left(\frac{R}{10 \text{ kpc}} \right)^2 M_{\odot} \text{ yr}^{-1}.$$

Here R is the radius of the area of the disk where a Fountain occurs, while f is the fraction of the disk covered by the flow. Although based on only a few datapoints, this shows that the flow rate obtained from UV line emission and from the observed IVCs is of the same order of magnitude, giving confidence in the Fountain picture and its associated numbers. However, in order to firm up these estimates, more and better distances are needed for IVCs, as are more observations of C IV (and O VI) emission.

7.2 Tidal Streams

The Magellanic Stream has long been identified as gas pulled out from the Small Magellanic Cloud (SMC), either by tidal forces or by ram-pressure stripping. Such an origin is strongly supported by measurements of the metallicity of $Z/Z_{\odot} \sim 0.3$ found for gas in both the leading and trailing parts of the Stream (Lu et al. 1998, Wakker et al. 2002, using NGC 3783, and Gibson et al. 2000, using Fairall 9). The implied presence of dust in the Stream also supports this idea (Richter et al. 2001c). Such a metallicity is consistent with the abundances for stars and gas in the LMC and SMC.

In theoretical models of tidal stripping (Gardiner and Noguchi 1996; Mastrogiro et al. 2005; Connors et al. 2006), the combined tidal force of the Milky Way and the LMC extracted the outer gaseous parts of the SMC during the previous perigalactic passage of the Magellanic Clouds, 1.5 Gyr ago. Affected by hydrodynamical interactions with the Milky Way's gaseous corona, decelerated and accelerated gas then formed a long stream, predicted to have distances of 50–100 kpc. During the most recent perigalacticon (~ 0.2 Gyr ago) more gas was extracted, resulting in the “Magellanic Bridge,” gas now located between the Clouds. This model is supported by the fair match between the observed and predicted locations and velocities of the

Stream, especially the so-called “Leading Arm.” Note that Putman et al. (1998) use this term for the short and small structure centered on $l = 320^\circ$, $b = -22^\circ$ (see [Fig. 12-1](#)). However, the main body of the Leading Arm extends to $l = 265^\circ - 290^\circ$, $b = +30^\circ$.

It is also possible that some of the gas in the Stream originated from ram pressure stripping, as the Magellanic Clouds move through the hot coronal medium surrounding the Milky Way. As discussed by Bregman (2004) and Bregman and Lloyd-Davies (2007), near the LMC the density of the coronal gas is at most about 10^{-4} cm^{-3} . Moving through this with a velocity of $\sim 300 \text{ km s}^{-1}$, the gas in the Clouds sees a ram pressure that is comparable to the gravitational force (see e.g., Moore and Davis 1994). Thus, gas in the LMC is safe, but gas in the outer parts of the SMC might be stripped. In the case of even smaller dwarf galaxies passing within 200 kpc from the Milky Way, this seems to have been the case, since none of the nearby dwarfs has much remaining gas (Grcevich and Putman 2009).

However, estimates of the proper motion of the MCs (Kallivayalil et al. 2006) and new cloud orbits (Besla et al. 2007) have thrown a wrench into this picture. These results would suggest much larger distances for the Stream ($\sim 150 \text{ kpc}$). In this case the Stream would consist of gas pulled out of the SMC by the LMC without the help of the Milky Way. A model calculation that predicts the distribution of Stream gas was made by Besla et al. (2010). It does a fair job of predicting the location and velocities of the main body of the Stream, as do the older tidal models. However, in this situation it is more difficult to form a Leading Arm that extends north of the Galactic plane. Besla et al. (2010) predict a modest amount of gas $10^\circ - 20^\circ$ ahead of the LMC. The tidal models in which the Stream has orbited the Milky Way more than once predict a Leading Arm that is much longer, which is required to explain the presence of gas with Magellanic metallicity seen toward NGC 3783, 50° from the LMC.

The Magellanic Stream is probably only the most well-defined and prominent example of the removal of gas from dwarf galaxies that orbit the Milky Way. Many other high-velocity clouds may have originated in this manner. However, in no other case has it been possible to make the link between a HVC and an identified existing or past dwarf. In the case of the cloud that has been named HVC 40–15+100, complex GCP, complex GP or “the Smith Cloud” by different authors, Lockman et al. (2008) model its orbit by assuming that the elongated appearance is due to tidal stretching, and interpreting the observed velocity gradient as a projection effect. Similarly, the elongated appearance of complexes A and C suggests that tidal stretching plays a role. However, there has not been a study that deduces the associated orbit, and no stellar counterparts have been identified for any of these HVCs.

One possible problem with identifying HVCs as tidal debris is that after the stripping takes place, the gas is no longer gravitationally bound and thus it will expand unless it is confined by hot coronal gas. However, in that case it will heat up and may evaporate, also disappearing from view in 21-cm observations.

7.3 Low-Metallicity Accretion

One of the most important processes traced by HVCs is the accretion of low-metallicity ($Z/Z_\odot \sim 0.1$) material onto the Galaxy, which dilutes the metals formed by stars. Such accretion is the theorists’ favorite method to solve the “G-dwarf problem” (see e.g., Pagel 1997), the

fact that the metallicity distribution of G-dwarfs (long-lived stars that have continually been born since the formation of the Milky Way disk) is narrower than that predicted by closed-box models. The theoretically favored models imply a present-day infall rate of about $0.4 M_{\odot} \text{ yr}^{-1}$ (Chiappini 2008).

Some of the HVCs provide evidence that such accretion is indeed ongoing. The metallicity of the second-largest HVC (complex C) has been measured as $Z/Z_{\odot} \sim 0.15$ (Wakker et al. 1999; Fox et al. 2004 and references therein). Furthermore, it has an (N/O) ratio that is about one fifth solar (this number includes ionization corrections). Thus, complex C consists of gas that has never been part of the Galaxy, but is about to accrete. Its distance is ~ 10 kpc (Wakker et al. 2007). Combined with its estimated vertical velocity ($50\text{--}150 \text{ km s}^{-1}$, depending on the location in the cloud) and total gas mass ($\sim 10^7 M_{\odot}$ in hydrogen and helium and, assuming a 50% hydrogen ionization fraction), this implies a mass accretion rate of $0.1\text{--}0.25 M_{\odot} \text{ yr}^{-1}$ in complex C alone. A number of other HVCs are also suspected of having a metallicity of $\sim 1/10$ th solar (e.g., complex A and the Anti-Center HVCs), although more evidence is needed. Tentative distance brackets on the order of 10 kpc exist for some of these clouds. If these metallicities and distances can be confirmed, they would combine to an additional mass accretion rate of about $0.05\text{--}0.1 M_{\odot} \text{ yr}^{-1}$. Further accounting for clouds in the population that are hidden by low-velocity gas (see [Sect. 2.5](#)), these rates suggest that the total accretion rate (summing both the neutral and the ionized phase) represented by low-metallicity HVCs is similar to the predicted amount.

The ultimate origin of this infalling gas is still unclear, however. An unknown fraction (possibly all) of the accreting gas may originally have been part of old tidal streams, similar to the Magellanic Stream. The Stream itself is a likely source of about $10^8 M_{\odot}$ of $Z/Z_{\odot} = 0.25$ gas over the next few Gyr. As long as the accreting gas has a metallicity on the order of 0.1 times solar, it makes little difference for the chemical evolution models whether the accreting gas originates in a tidal stream or is more pristine. However, the amount of gas available from the known population of dwarf galaxies near the Milky Way (and near M 31) is several orders of magnitude below the required infall rate (Grcevich and Putman 2009), so that a fully tidal origin seems unlikely. Thus, a different mechanism is needed. Three mechanisms have been proposed (see Dekel and Birnboim 2006), which lead to subtly different predictions for the behavior of the HVC population.

7.3.1 Fragmentation in a Hot Halo

The first mechanism for the origin of HVCs is that they are the observable result of the fragmentation of a hot corona as it cools. This kind of model was originally proposed by Oort (1970), who suggested that it might take a long time before all the gas in the Milky Way's gravitational sphere of influence would make it to the disk. In a more hydrodynamically oriented version of this idea (White and Rees 1978), intergalactic gas that falls into dark matter potential wells heats to the virial temperature ($T = 10^6 (v_{\text{max}}/163)^2 \text{ K}$), and the denser gas at small radii then cools and accretes. This reduces the pressure support of the more distant gas, which falls in, gets denser, and cools in turn. Following ideas of Mo and Miralda-Escude (1996), Maller and Bullock (2004) use this picture, but work out the consequences of dropping the assumption that the cooling is uniform. Instead they assume that clouds form by fragmentation, and then derive the typical parameters of such clouds. Using the same notation as in [Sect. 7.1](#), the energy loss per unit mass is again given by $Q = \Lambda n_H / m_a$, but now Q is given by $(3/2 k T_h) / (m_p t_f)$, with t_f

the time since the halo formed. This gives a characteristic “cooling density,” ρ_c , the mass density above which the gas can cool after time t_f :

$$\rho_c = \frac{3 m_a^2 k T_h}{2 m_p^2 t_f \Lambda(T, Z)},$$

where T_h is the temperature of the hot halo gas. Note that here it is assumed that $\Lambda(T, Z)$ is proportional to the metallicity, Z , of the gas. For a halo in which gas has a metallicity $Z = 0.1$ solar, and taking $t_f = 8$ Gyr, the number density n_H then is $\sim 1.5 \times 10^{-4} \text{ cm}^{-3}$. Now, when a mostly neutral cloud forms (having temperature $T_w \sim 10^4$ K, and density ρ_w), it will end up in pressure equilibrium with the hot gas. Defining the average halo pressure as \bar{P}_h , then $P_w = \bar{P}_h$ for the typical cloud that forms. Maller and Bullock (2004) show that $\bar{P}_h \sim 2.7 P_c = \eta_P P_c$, with P_c the pressure corresponding to ρ_c . This implies $\rho_w = \rho_c (\eta_P T_h / T_w)$, so that:

$$r_{\text{cl}} = \left(\frac{3 M_{\text{cl}}}{4\pi \rho_w} \right)^{1/3} \sim 1050 \left(\frac{M}{5 \times 10^6 M_\odot} \right)^{1/3} \left(\frac{T_h}{10^6 \text{ K}} \right)^{-1/2} \left(\frac{T_w 10^6}{10^4 T_h} \Lambda_Z \frac{t_f}{10^8 \text{ Gyr}} \right)^{1/3} \text{ pc},$$

where r_{cl} is the cloud’s radius, M is the cloud mass, $\Lambda_Z = \Lambda(10^6, Z) / \Lambda(T, Z)$ is the value of the cooling function relative to that at $T = 10^6$ K, and metallicity $Z = 0.1$ solar. The only thing left to determine is the typical cloud mass. Maller and Bullock (2004) present arguments based on the limits set by different physical processes (evaporation, conduction, thermal instabilities, drag, self-gravity) and conclude that for a Milky Way–sized halo, $M_{\text{cl}} \sim 5 \times 10^6 M_\odot$. The infall of these clouds is then driven by drag and cloud–cloud collisions, from which they derive the typical infall time and thus the growth rate of a galaxy.

This approach predicts the presence of clouds with typical diameter about $2.4 (D/50 \text{ kpc})$, and (assuming 50% ionization) a typical H I column density $4 \times 10^{19} \text{ cm}^{-2}$, which is just a little bit larger and brighter than the typical properties of the observed population of small HVCs. However, the model also predicts a total mass in cold clouds of about $2 \times 10^{10} M_\odot$, which would imply the presence of 4,000 HVC-like clouds in the halo. This is an order of magnitude larger than the number that is observed. Reconciling this with the observed population might be done by assuming a much smaller neutral fraction ($< 5\%$), making the implied H I column density smaller, so that most clouds would go undetected in 21-cm H I emission.

7.3.2 Cold, Filamentary Streams

Kereš and Hernquist (2009) used simulations to confirm an alternative way to produce cool clouds embedded in hot coronal gas as a natural consequence of structure and galaxy formation in the universe. Their model consists of a cosmological simulation of structure formation caused by dark matter in a $10h^{-1} \text{ Mpc}$ box. At $z = 99$ they populate a Milky Way–sized box ($200h^{-1} \text{ kpc}$ on the side) with gas particles. Following the hydrodynamic and thermodynamic state of this gas (including heating, average cooling, and a simple prescription for ionization by UV photons, but excluding outflows from the galaxies in the box), they then find that some cool neutral clouds that lose angular momentum and accrete are formed in the “hot mode” manner described above. However, some of the gas never gets hotter than 10^5 K before it reaches the galaxy disk. Furthermore, this gas streams in mostly along the large-scale filaments created by the dark matter, causing an asymmetric inflow, unlike what is the case for the hot-mode accretion. According to Kereš et al. (2009), for dark matter halos below $5 \times 10^{10} M_\odot$ all accretion

follows the cold-mode route, while in dark matter halos above $3 \times 10^{11} M_{\odot}$ (galaxy mass above $3 \times 10^{10} M_{\odot}$) hot mode accretion is more important.

The model predicts the existence of 100–1,000 clouds with $T < 30,000$ K and $n > 0.004 \text{ cm}^{-3}$ at distances up to about 50 kpc, with masses of 10^6 to a few $10^7 M_{\odot}$ (although the lower limit is probably due to the resolution of the simulation), accreting at a rate of $\sim 1 M_{\odot} \text{ yr}^{-1}$. At $z \sim 0$ about 10% of the gaseous halo mass is in the cold phase, which is in a dynamical equilibrium between cloud formation, destruction, and accretion. Clearly, the properties of this cloud population are similar to those of the observed HVCs, except for the density, which is observed to be on the order of 0.1 cm^{-3} for the larger clouds about 10 kpc from the plane, although estimates suggest a value more like 0.01 cm^{-3} for small (likely to be more distant) clouds. Further, in this model the infalling clouds have no metals.

Clearly, current understanding of the formation of clouds in a hot corona that surrounds galaxies predicts that HVC-like objects can form and accrete. What is not clear yet is which of the two modes is most compatible with the detailed characteristics of the HVCs, that is, their distribution with longitude and latitude, their velocities, their metallicities, their internal structure, density, etc. Furthermore, the physical processes used in the models are either only approximated analytically, or follow simple prescriptions for the microphysics to predict the conditions in $10^6 M_{\odot}$ gas “particles,” while the effects of UV radiation and feedback flows are only partially dealt with. Finally, in these models the state of the gas is not followed at temperatures below 10^4 K.

7.3.3 Sweeping up of Coronal Gas by Fountain Flows

In a series of papers, Fraternali and collaborators have considered the effects of the interaction of Fountain clouds with hot coronal gas. Fraternali and Binney (2006) pointed out that a continuous-outflow model of Fountain clouds can reproduce the z -distribution of extraplanar H I that is observed in the nearby spirals NGC 891 and NGC 2403 (see Sects. 6.2 and 6.3). However, this model fails to reproduce the observed vertical gradient in the rotation velocity and it also predicts a radial outflow rather than the observed radial inflow.

Fraternali and Binney (2008), Marinacci et al. (2010), and Marasco et al. (2011) show that these kinematic failures can be resolved if Fountain clouds sweep up and cause cooling in hot coronal gas. The amount of gas stripped from a Fountain cloud might be similar to the amount of cooled coronal gas condensed in the cloud’s wake. The result would be a disk accretion rate, valid also for the Milky Way Galaxy, of about two solar masses per year, with half of that fresh low-metallicity gas.

This model also can be fit to the LAB data for the Milky Way. Marasco and Fraternali (2011) find that the data can be reproduced using a model with a scale height of halo H I (1.6 kpc), a vertical rotation gradient of $15 \text{ km s}^{-1} \text{ kpc}^{-1}$ and a radial inflow of 30 km s^{-1} .

7.4 Summary of Origins

High-velocity clouds (broadly defined) appear to have multiple origins.

- (1) Some clouds with velocities that deviate by up to about 80 km s^{-1} from differential galactic rotation have metallicities that are near solar (the IV Arch, the LLIV Arch,

and the PP Arch) and distances on the order of 1 kpc or less. Some clouds are moving away from the disk at $50\text{--}80\text{ km s}^{-1}$ (cloud g1) and have similar amounts of ionized and neutral hydrogen. These clouds fit the predicted properties of a Galactic Fountain return flow and outflow. From their masses, distances, and velocities it is possible to derive the circulation rate of gas between the Milky Way disk and halo. Preliminary estimates give values on the order of $4.5 M_{\odot} \text{ yr}^{-1}$. This implies that all of the ISM ($\sim 5 \times 10^9 M_{\odot}$) circulates through the lower halo on timescales of a Gyr.

- (2) Other HVCs clearly are connected to the Magellanic Clouds, in particular the Magellanic Stream, which originates from the Clouds. The Stream gas has metallicities similar to those in the SMC. It most likely has a tidal origin, although ram pressure effects may have contributed to its formation, and certainly they are contributing to its current location and velocities. A number of HVCs are not in the main body of the Stream, but based on their location, velocities, internal properties, and metallicities they appear related to the processes that formed the Stream; this is especially true for the clouds with velocities $>200\text{ km s}^{-1}$ in the region $l = 260^{\circ}\text{--}300^{\circ}$, $b = -20^{\circ}$ to $+30^{\circ}$. The discovery of O VI absorption associated with the Stream shows that it is embedded in hot ($>10^6\text{ K}$) coronal gas. Note that this implied corona has a different origin than the lower-temperature corona that is formed at heights a few kpc above the Galactic disk by the escape of hot gas from superbubbles and supernovae.
- (3) The properties of the two large complexes in the northern sky (complex A and C) suggest that they represent the last stages of the infall of low-metallicity material from intergalactic space – they are elongated, suggesting tidal stretching, and in the case of complex C the measured metallicity is ~ 0.15 times solar, the D/H ratio is high, and the N/O ratio is subsolar. Other clouds, such as many (or all) of the ones in the Anti-Center region, may also fit into this category. The total infall rate of baryons associated with these objects is on the order of $0.2\text{--}0.4 M_{\odot} \text{ yr}^{-1}$ (assuming equal amounts of neutral and ionized gas in these clouds).
- (4) Finally, there are many smaller HVCs, which have a size spectrum that is a power law, and whose velocities and distribution on the sky can be understood in a model in which they have distances up to 80 kpc and they mostly orbit the Milky Way, but with random velocities that have a dispersion of 50 km s^{-1} and a net infall velocity of 50 km s^{-1} for the population. For some of these clouds distances of 10–20 kpc have been found. Models of structure formation in the universe predict similar populations of clouds, but do not yet correctly predict the detailed cloud properties.

The challenges for the future include the following: (a) Determining for each HVC into which of these possible origins it fits, so that a proper determination of the circulation and accretion rates becomes possible. (b) Comparing the observations of the locations and velocities of the HVCs with the predictions for the formation of Galactic Fountain clouds, the Magellanic Stream and condensations in hot coronas. (c) Comparing the observed physical conditions (density, temperature, pressure, ionization fraction, interface properties) of the clouds with the predictions of the different models. (d) Finding HVC analogs around external galaxies, to allow generalizing the understanding of such objects and determining correlations with galaxy properties.

Cross-References

- ▶ [Astrophysics of Galactic Charged Cosmic Rays](#)
- ▶ [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- ▶ [Dynamics of Disks and Warps](#)
- ▶ [Galactic Distance Scales](#)
- ▶ [Gamma-Ray Emission of Supernova Remnants and the Origin of Galactic Cosmic Rays](#)
- ▶ [History of Dark Matter in Galaxies](#)
- ▶ [Magnetic Fields in Galaxies](#)
- ▶ [Mass Distribution and Rotation Curve in the Galaxy](#)

References

- Adams, W. S. 1949, *ApJ*, 109, 354
- Allen, R. J., Goss, W. M., & van Woerden, H. 1973, *A&A*, 29, 447
- Arnal, E. M., Bajaja, E., Larrarte, J. J., Morras, R., & Pöppel, W. G. L. 2000, *A&AS*, 142, 35
- Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, *ARA&A*, 47, 481
- Bajaja, E., Cappa de Nicolau, C. E., Cersosimo, J. C., Martin, M. C., Loiseau, N., Morras, R., Olano, C. A., & Pöppel, W. G. L. 1985, *ApJS*, 58, 143
- Barger, K. A., Haffner, L. M., Hill, A. S., Wakker, B. P., Madsen, G. J., & Duncan, A. K. 2012, *ApJ* (submitted)
- Beers, T. C., Wilhelm, R., Doinidis, S. P., & Mattson, C. J. 1996, *ApJS*, 103, 433
- Begelman, M. C., & Fabian, A. C. 1990, *MNRAS*, 244, 26
- Benjamin, R. A. 2004, in *High-Velocity Clouds, Astrophysics and Space Science Library*, Vol. 312 (Dordrecht: Kluwer), 371
- Benjamin, R. A., & Danly, L. 1997, *ApJ*, 481, 764
- Bertoldi, F., & McKee, C. F. 1992, *ApJ*, 395, 140
- Besla, G., Kallivayalil, N., Hernquist, L., Robertson, B., Cox, T. J., van der Marel, R. P., & Alcock, C. 2007, *ApJ*, 668, 949
- Besla, G., Kallivayalil, N., Hernquist, L., van der Marel, R. P., Cox, T. J., & Kerés, D. 2010, *ApJ*, 721, L97
- Bland-Hawthorn, J., & Maloney, P. R. 1999, *ApJ*, 510, L33 (erratum 2001, *ApJ*, 553, L231)
- Bland-Hawthorn, J., Veilleux, S., Cecil, G. N., Putman, M. E., Gibson, B. K., & Maloney, P. R. 1998, *MNRAS*, 299, 611
- Blitz, L., Spergel, D., Teuben, P., Hartmann, D., & Burton, W. B. 1999, *ApJ*, 514, 818
- Boomsma, R., Oosterloo, T. A., Fraternali, F., van der Hulst, J. M., & Sancisi, R. 2008, *A&A*, 490, 555
- Borkowski, K. J., Balbus, S. A., & Frstrom, C. C. 1990, *ApJ*, 355, 501
- Boulanger, F., Abergel, A., Bernard, J. P., Burton, W. B., Désert, F.-X., Hartmann, D., Lagache, G., & Puget, J.-L. 1996, *A&A*, 312, 256
- Bowen, D. V., Jenkins, E. B., Tripp, T. M., Sembach, K. R., Savage, B. D., Moos, H. W., Oegerle, W. R., Friedman, S. D., Gry, C., Kruk, J. W., Murphy, E., Sankrit, R., Shull, J. M., Sonneborn, G., & York, D. G. 2008, *ApJS*, 176, 59
- Braun, R., & Burton, W. B. 1999, *A&A*, 341, 437
- Braun, R., & Thilker, D. A. 2004, *A&A*, 417, 421
- Breitschwerdt, D., McKenzie, J. F., & Völk, H. J. 1993, *A&A*, 269, 54
- Bregman, J. N. 1980, *ApJ*, 236, 577
- Bregman, J. N. 2004, in *High-Velocity Clouds, Astrophysics and Space Science Library*, Vol. 312, (Dordrecht: Kluwer), 341
- Bregman, J. N., & Lloyd-Davies, E. J. 2007, *ApJ*, 669, 990
- Brinks, E., & Bajaja, E. 1986, *A&A*, 169, 14
- Brinks, E., & Burton, W. B. 1984, *A&A*, 141, 195
- Brown, W. R., Geller, J. M., Kenyon, S. J., Beers, T. C., Kurtz, M. J., & Roll, J. B. 2004, *AJ*, 127, 1555
- Brüns, C., Kerp, J., & Pagels, A. 2001, *A&A*, 370, L26
- Burkhart, B., Falceta-Gonçalves, D., Kowal, G., & Lazarian, A. 2009, *ApJ*, 693, 250
- Chiappini, C. 2008, *ASP Conf Ser*, 396, 113
- Collins, J. A., Shull, J. M., & Giroux, M. L. 2003, *ApJ*, 585, 336
- Collins, J. A., Shull, J. M., & Giroux, M. L. 2007, *ApJ*, 657, 271
- Connors, T. W., Kawata, D., & Gibson, B. K. 2006, *MNRAS*, 371, 108
- Cutri, R. M., et al. 2003, 2MASS All Sky Catalog of Point Sources (VizieR Online Data Catalog #2246)
- Danly, L., Albert, C. E., & Kuntz, K. D. 1993, *ApJ*, 416, L29

- Davies, R. D., Buhl, D., & Jafolla, J. 1976, *A&AS*, 23, 181
- de Avillez, M. A., & Breitschwerdt, D. 2005, *A&A*, 436, 585
- de Heij, V., Braun, R., & Burton, W. B. 2002a, *A&A*, 391, 67
- de Heij, V., Braun, R., & Burton, W. B. 2002b, *A&A*, 392, 417
- Dehnen, W., & Binney, J. 1998, *MNRAS*, 294, 429
- Dekel, A., & Birnboim, Y. 2006, *MNRAS*, 368, 2
- Désert, F.-X., Bazell, D., & Blitz, L. 1990, *ApJ*, 355, L51
- Deul, E. R., & den Hartog, R. H. 1990, *A&A*, 229, 362
- Dopita, M. A., & Sutherland, R. S. 1996, *ApJS*, 102, 161
- Espreate, J., Cantó, J., & Franco, J. 2002, *ApJ*, 575, 194
- Esquivel, A., Benjamin, R. A., Lazarian, A., Cho, J., & Leitner, S. N. 2006, *ApJ*, 648, 1043
- Everett, J. E., Zweibel, E. G., Benjamin, R. A., McCammon, D., Rocks, L., & Gallagher, J. S. 2008, *ApJ*, 674, 258
- Ferland, G. 1996, Hazy, a brief introduction to CLOUDY, Univ. Kentucky Ph. Dept. Report
- Ferrara, A., & Field, G. B. 1994, *ApJ*, 423, 665
- Fitzpatrick, E. L., & Spitzer, L. 1997, *ApJ*, 475, 623
- Fox, A. J., Savage, B. D., Wakker, B. P., Richter, P., Sembach, K. R., & Tripp, T. M. 2004, *ApJ*, 602, 738
- Fox, A. J., Savage, B. D., & Wakker, B. P. 2006, *ApJS*, 165, 229
- Fraternali, F. 2010, *AIP Conf*, 1240, 135
- Fraternali, F., & Binney, J. 2006, *MNRAS*, 366, 449
- Fraternali, F., & Binney, J. 2008, 386, 935
- Fraternali, F., van Moorsel, G., Sancisi, R., & Oosterloo, T. A. 2002a, *AJ*, 123, 3124
- Fraternali, F., Cappi, M., Sancisi, R., & Oosterloo, T. A. 2002b, *ApJ*, 578, 109
- Fraternali, F., Oosterloo, T. A., & Sancisi, R. 2004, *A&A*, 424, 485
- Fraternali, F., Oosterloo, T. A., Sancisi, R., & Swaters, R. A. 2005, *ASP Conf*, 331, 239
- Fujimoto, M., & Sofue, Y. 1976, *A&A*, 47, 263
- Gardiner, L. T., & Noguchi, M. 1996, *MNRAS*, 278, 191
- Gibson, B. K., Giroux, M. L., Penton, S. V., Putman, M. E., Stocke, J. T., & Shull, J. M. 2000, *AJ*, 120, 1830
- Gibson, B. K., Giroux, M. L., Penton, S. V., Stocke, J. T., Shull, J. M., & Tumlinson, J. 2001, *AJ*, 122, 3280
- Giovanelli, R. 1980, *AJ*, 85, 1155
- Giovanelli, R., & Haynes, M. P. 1977, *A&A*, 54, 909
- Giovanelli, R., Verschuur, G. L., & Cram, T. R. 1973, *A&AS*, 12, 209
- Girardi, L., Bertelli, G., Bressan, A., Chiosi, C., Groenewegen, M. A. T., Marigo, P., Salasnich, B., & Weiss, A. 2002, *A&A*, 391, 195
- Gnat, O., & Sternberg, A. 2007, *ApJS*, 168, 213
- Gnat, O., Sternberg, A., & McKee, C. F. 2009, *ApJ*, 718, 1315
- Grcevich, J., & Putman, M. E. 2009, *ApJ*, 696, 385
- Green, R. F., Schmidt, M., & Liebert, J. 1986, *ApJS*, 61, 305
- Haffner, L. M. 2005, in *ASP Conf. Proc.*, *Extraplanar Gas*, Vol. 331, ed. R. Braun, 25, Kluwer
- Haffner, L. M., Reynolds, R. J., & Tufte, S. L. 2001, *ApJ*, 556, L33
- Haffner, L. M., Reynolds, R. J., Tufte, S. L., Madsen, G. J., Jaehnig, K. P., & Percival, J. W. 2003, *ApJS*, 149, 405
- Hartmann, D., & Burton, W. B. 1997, *Atlas of Galactic Neutral Hydrogen* (Cambridge: Cambridge University Press)
- Heald, G. H., Rand, R. J., Benjamin, R. A., Collins, J. A., & Bland-Hawthorn, J. 2006a, *ApJ*, 636, 181
- Heald, G. H., Rand, R. J., Benjamin, R. A., & Bershad, M. A. 2006b, *ApJ*, 647, 1018
- Hill, A. S., Haffner, L. M., & Reynolds, R. J. 2009, *ApJ*, 703, 1832
- Houck, J. C., & Bregman, J. N. 1990, *ApJ*, 352, 506
- Hulsbosch, A. N. M. 1975, *A&A*, 40, 1
- Hulsbosch, A. N. M., & Wakker, B. P. 1988, *A&AS*, 75, 191
- Ibata, R., Irwin, M., Lewis, G., Ferguson, A. N. M., & Tanvir, N. 2001, *Nature*, 412, 49
- Ivezić, Ž., Vivas, A. K., Lupton, R. H., & Zinn, R. 2005, *AJ*, 129, 1096
- Kahn, F. D. 1976, *A&A*, 50, 145
- Kahn, F. D. 1981, in *Investigating the Universe: Papers presented to Zdenek Kopal on the Occasion of his Retirement, September 1981* (Dordrecht: Reidel), 1
- Kalberla, P. M. W., Burton, W. B., Hartmann, D., Arnal, E. M., Bajaja, E., Morras, R., & Pöppel, W. G. L. 2005, *A&A*, 440, 775
- Kalberla, P. M. W., McClure-Griffiths, N. M., Pisano, D. J., Calabretta, M. R., Ford, H. A., Lockman, F. J., Staveley-Smith, L., Kerp, J., Winkel, B., Murphy, T., & Newton-McGee, K. 2010, *A&A*, 521, A17
- Kallivayalil, N., van der Marel, R. P., & Alcock, C. 2006, *ApJ*, 652, 1213
- Kamphuis, J. J., Sancisi, R., & van der Hulst, J. M. 1991, *A&A*, 244, L29
- Kamphuis, J. J., & Sancisi, R. 1993, *A&A*, 273, L31
- Kereš, D., & Hernquist, L. 2009, *ApJ*, 700, L1
- Kereš, D., Katz, N., Fardal, M., Davé, R., & Weinberg, D. H. 2009, *MNRAS*, 395, 160

- Klypin, A., Kravtsov, A. V., Valenzuela, O., & Prada, F. 1999, *ApJ*, 522, 82
- Kravtsov, A., *AdAstr* 2010, E8, *Advances in Astronomy*, Volume 2010, Article ID 281913
- Kuijken, K., & Gilmore, G. 1991, *ApJ*, 367, L9
- Kukarkin, B. V., et al. 1970, *General Catalogue of Variable Stars*, Vols. I and II (3rd ed.; Moscow: Sternberg Institute)
- Kuntz, K. D., & Danly, L. 1996, *ApJ*, 457, 703
- Kwak, K., & Shelton, R. L. 2010, *ApJ*, 719, 523
- Lazarian, A., & Pogosyan, D. 2000, *ApJ*, 537, 720
- Lee, S. W., Irwin, J. A., Dettmar, R. J., Cunningham, C. T., Golla, G., & Wang, Q. D. 2001, *A&A*, 377, 759
- Lehner, N., & Howk, J. C. 2011, *Science*, 334, 955
- Lin, D. N. C., & Lynden-Bell, D. 1982, *MNRAS*, 198, 707
- Lockman, F. J. 2003, *ApJ*, 591, L33
- Lockman, F. J., Benjamin, R. A., Heroux, A. J., & Langston, G. I. 2008, *ApJ*, 679, L21
- Lu, L., Savage, B. D., Sembach, K. R., Wakker, B. P., Sargent, W. L. W., & Oosterloo, T. A. 1998, *AJ*, 115, 162
- Mac Low, M.-M., McCray, R., & Norman, M. L. 1989, *ApJ*, 337, 141
- Madsen, G. J., Reynolds, R. J., & Haffner, L. M. 2006, *ApJ*, 652, 401
- Magnani, L., & Smith, A. J. 2010, *ApJ*, 722, 1685
- Maller, A. H., & Bullock, J. S. 2004, *MNRAS*, 355, 694
- Maloney, P. 1993, *ApJ*, 414, 41
- Maloney, P. R., & Putman, M. E. 2003, *ApJ*, 589, 270
- Marasco, A., & Fraternali, F. 2011, *A & A*, 525, A134
- Marasco, A., Fraternali, F., & Binney, J. 2011, *MNRAS*, 419, 1107
- Marinacci, F., Binney, J., Fraternali, F., Nipoti, C., Ciotti, L., & Londrill, P. 2010, *MNRAS*, 404, 1464
- Martin, C., & Bowyer, S. 1990, *ApJ*, 350, 242
- Mastropietro, C., Moore, B., Mayer, L., Wadsley, J., & Stadel, J. 2005, *MNRAS*, 363, 509
- Mathis, J. S., Mezger, P. G., & Panagia, N. 1983, *A&A*, 128, 212
- Matthews, L. D., & Wood, K. 2003, *ApJ*, 593, 721
- McCammon, D., & Sanders, W. T. 1990, *ARA&A*, 28, 657
- McConnachie, A. W. et al. 2009, *Nature*, 461, 66
- Mo, H. J., & Miralda-Escude, J. 1996, *ApJ*, 469, 589
- Moore, B., & Davis, M. 1994, *MNRAS*, 270, 209
- Moore, B., Ghigna, S., Governato, F., Lake, G., Quinn, T., Stadel, J., & Tozzi, P. 1999, *ApJ*, 524, L19
- Morras, R., Bajaja, E., Arnal, E. M., & Pöppel, W. G. L. 2000, *A&AS*, 142, 25
- Morton, D. C. 2003, *ApJS*, 149, 205
- Muller, C. A., Oort, J. H., & Raimond, E. 1963, *C R Acad Sci Paris*, 257, 1661
- Münch, G. 1952, *PASP*, 64, 312
- Murphy, E. M., Lockman, F. J., & Savage, B. D. 1995, *ApJ*, 447, 642
- Oort, J. H. 1966, *BAN*, 18, 421
- Oort, J. H. 1970, *A&A*, 7, 381
- Oosterloo, T. A. 2004, in *High-Velocity Clouds*, *ASSL*, 312, 125
- Oosterloo, T. A., Fraternali, F., & Sancisi, R. 2007, *AJ*, 134, 1019
- Osterbrock, D. E. 1989, *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (Mill Valley: University Science Books)
- Pagel, B. 1997, *Nucleosynthesis and Chemical Evolution of Galaxies* (Cambridge: Cambridge University Press)
- Peek, J. E. G., Heiles, C., Douglas, K. A., Lee, M.-Y., Grcevich, J., Stanimirovic, S., Putman, M. E., Korpela, E. J., Gibson, S. J., Begum, A., Saul, D., Tobishaw, T., & Krco, M. 2011, *ApJS*, 194, 20
- Pildis, R. A., Bregman, J. N., & Schombert, J. M. 1994, *ApJ*, 423, 190
- Pisano, D. J., Barnes, D. G., Gibson, B. K., Staveley-Smith, L., Freeman, K. C., & Kilborn, V. A. 2007, *ApJ*, 662, 959
- Preston, G. W., Schectman, S. A., & Beers, T. C. 1991, *ApJS*, 76, 1001
- Putman, M. E., Gibson, B. K., Staveley-Smith, L., et al. 1998, *Nature*, 394, 752
- Putman, M. E., Bland-Hawthorn, J., Veilleux, S., Gibson, B. K., Freeman, K. C., & Maloney, P. R. 2003, *ApJ*, 597, 948
- Putman, M. E., Peek, J. E. G., Muratov, A., Gnedin, O. Y., Hsu, W., Douglas, K. A., Heiles, C., Stanimirovic, S., Korpela, E. J., & Gibson, S. J. 2009, *ApJ*, 703, 1486
- Rand, R. J. 2000, *ApJ*, 537, L13
- Rand, R. J., & van der Hulst, J. M. 1993a, *AJ*, 105, 2098
- Rand, R. J., & van der Hulst, J. M. 1993b, *AJ*, 107, 392
- Reach, W. T., Koo, B.-C., & Heiles, C. 1994, *ApJ*, 429, 672
- Richter, P., Savage, B. D., Wakker, B. P., Sembach, K. R., & Kalberla, P. M. W. 2001a, *ApJ*, 549, 281
- Richter, P., Sembach, K. R., Wakker, B. P., Savage, B. D., Tripp, T. M., Murphy, E. M., Kalberla, P. M. W., & Jenkins, E. B. 2001b, *ApJ*, 559, 318
- Richter, P., Sembach, K. R., Wakker, B. P., & Savage, B. D. 2001c, *ApJ*, 562, L181
- Richter, P., Wakker, B. P., Savage, B. D., & Sembach, K. R. 2003, *ApJ*, 586, 230
- Rosen, A., & Bregman, J. N. 1995, *ApJ*, 440, 634
- Ryans, R. S. I., Keenan, F. P., Sembach, K. R., & Davies, R. D. 1997, *MNRAS*, 289, 83
- Sancisi, R., & Allen, R. J. 1979, *A&A*, 74, 73
- Sancisi, R., Fraternali, F., Oosterloo, T. A., & van der Hulst, J. M. 2008, *A&ARv*, 15, 189

- Savage, B. D., & Sembach, K. R. 1996, *ARA&A*, 34, 279
- Savage, B. D., & Wakker, B. P. 2009, *ApJ*, 702, 1472
- Savage, B. D., Meade, M. R., & Sembach, K. R. 2001, *ApJS*, 136, 631
- Savage, B. D., Sembach, K. R., Wakker, B. P., Richter, P., Meade, M., Jenkins, E. B., Shull, J. M., Moos, H. W., & Sonneborn, G. 2003, *ApJS*, 146, 125
- Schaap, W. E., Sancisi, R., & Swaters, R. A. 2000, *A&A*, 356, L49
- Schlüter, A., Schmidt, H., & Stumpff, P. 1953, *Zeitschrift für Astrophysik*, 33, 194
- Schwarz, U. J., & Oort, J. H. 1981, *A&A*, 101, 305
- Schwarz, U. J., & Wakker, B. P. 2004, in *High-Velocity Clouds, Astrophysics and Space Science Library*, Vol. 312 (Dordrecht: Kluwer), 145
- Schwarz, U. J., Wakker, B. P., & van Woerden, H. 1995, *A&A*, 302, 364
- Sembach, K. R., & Savage, B. D. 1992, *ApJS*, 83, 147
- Sembach, K. R., Savage, B. D., & Massa, D. 1991, *ApJ*, 372, 81
- Sembach, K., Savage, B. D., Lu, L., & Murphy, E. M. 1995, *ApJ*, 451, 616
- Sembach, K., Savage, B. D., Lu, L., & Murphy, E. M. 1999, *ApJ*, 515, 108
- Sembach, K. R., Howk, J. C., Savage, B. D., & Shull, J. M. 2001, *AJ*, 121, 992
- Sembach, K. R., Gibson, B. K., Fenner, Y., & Putman, M. E. 2002, *ApJ*, 572, 178
- Sembach, K. R., Wakker, B. P., Savage, B. D., Richter, P., Meade, M., Shull, J. M., Jenkins, E. B., Sonneborn, G., & Moos, H. W. 2003, *ApJS*, 146, 165
- Sembach, K. R., Wakker, B. P., Tripp, T. M., Richter, P., Kruk, J. W., Blair, W. P., Moos, H. W., Savage, B. D., Shull, J. M., York, D. G., Sonneborn, G., Hébrard, G., Ferlet, R., Vidal-Madjar, A., Friedman, S. D., & Jenkins, E. B. 2004, *ApJS*, 150, 387
- Shapiro, P. R., & Benjamin, R. 1991, *PASP*, 103, 923
- Shapiro, P. R., & Field, G. B. 1976, *ApJ*, 205, 762
- Shelton, R. 1998, *ApJ*, 504, 785
- Shelton, R. L., Sallmen, S. M., & Jenkins, E. B. 2007, *ApJ*, 659, 365
- Shull, J. M., Stevans, M., Danforth, C., Penton, S. V., Lockman, F. J., & Arav, N. 2011, *ApJ*, 739, 105
- Sirko, E., Goodman, J., Knapp, G. R., Brinkmann, J., Ivezić, Ž., Knerr, E. J., Schlegel, D., Schneider, D. P., & York, D. G. 2004, *AJ*, 127, 899
- Slavin, J. D., Shull, J. M., & Begelman, M. C. 1993, *ApJ*, 407, 83
- Smoker, J. V., Hunter, I., Kalberla, P. M. W., Keenan, F. P., Morras, R., Hanuschik, R., Thompson, H. M. A., Silva, D., Bajaja, E., Pöppel, W. G. L., & Arnal, M. 2007, *MNRAS*, 378, 947
- Snowden, S. L., Egger, R., Finkbeiner, D. P., Freyberg, M. J., & Plucinsky, P. P. 1998, *ApJ*, 493, 715
- Spitzer, L. 1956, *ApJ*, 124, 20
- Spitzer, L. 1998, *Physical Processes in the Interstellar Medium* (New York: Wiley-VCH)
- Stocke, J. T., Penton, S. V., Danforth, C. W., Shull, J. M., Tumlinson, J., & McClintock, K. M. 2006, *ApJ*, 641, 217
- Stoppelenburg, P. S., Schwarz, U. J., & van Woerden, H. 1998, *A&A*, 338, 200
- Sutherland, R. S., & Dopita, M. A. 1993, *ApJS*, 88, 253
- Swaters, R. A., Sancisi, R., & van der Hulst, J. M. 1997, *ApJ*, 491, 140
- Thilker, D. A., Braun, R., & Walterbos, R. A. M. 1998, *A&A*, 332, 429
- Thilker, D. A., Braun, R., Walterbos, R. A. M., Corbelli, E., Lockman, F. J., Murphy, E., & Madalena, R. 2004, *ApJ*, 601, L39
- Thom, C., Putman, M. E., Gibson, B. K., Christlieb, N., Flynn, C., Beers, T. C., Wilhelm, R., & Lee, Y. S. 2006, *ApJ*, 637, L97
- Thom, C., Peek, J. E. G., Putman, M. E., Heiles, C., Peek, K. M. G., & Wilhelm, R. 2008, *ApJ*, 684, 364
- Tripp, T. M., Wakker, B. P., Jenkins, E. B., Bowers, C. W., Danks, A. C., Green, R. F., Heap, S. R., Joseph, C. L., Kaiser, M. E., Linsky, J. L., & Woodgate, B. E. 2003, *AJ*, 125, 3122
- Tufte, S. L., Reynolds, R. J., & Haffner, L. M. 1998, *ApJ*, 504, 773
- van Dishoeck, E. F., & Black, J. H. 1986, *ApJS*, 62, 109
- van der Hulst, J. M., & Sancisi, R. 1988, *AJ*, 95, 1354
- van der Hulst, J. M., van Albada, T. S., & Sancisi, R. 2001, in *Gas and Galaxy Evolution*, ASPC 240 (San Francisco, CA: Astronomical Society of the Pacific), 451
- van Woerden, H., & Wakker, B. P. 2004, in *High-Velocity Clouds, Astrophysics and Space Science Library*, Vol. 312 (Dordrecht: Kluwer), 195
- van Woerden, H., Schwarz, U. J., Peletier, R. F., Wakker, B. P., & Kalberla, P. M. W. 1999a, *Nature*, 400, 138
- van Woerden, H., Peletier, R. D., Schwarz, U. J., Wakker, B. P., & Kalberla, P. M. W. 1999b, in *ASP Conf. Ser. 166, Stromlo Workshop on High-Velocity Clouds*, eds. B. K. Gibson, & M. E. Putman (San Francisco, CA: ASP), 1
- van Woerden, H., Wakker, B. P., Schwarz, U. J., & de Boer, K. S., eds. 2004, *High-Velocity Clouds, Astrophysics and Space Science Library*, Vol. 312 (Dordrecht: Kluwer)
- Verheijen, M. A. W., & Sancisi, R. 2001, *A&A*, 370, 765

- Wakker, B. P. 1991, *A&A*, 250, 499
- Wakker, B. P. 2001, *ApJS*, 136, 463
- Wakker, B. P. 2004, in *High-Velocity Clouds, Astrophysics and Space Science Library*, Vol. 312 (Dordrecht: Kluwer), 25
- Wakker, B. P. 2006, *ApJS*, 163, 282
- Wakker, B. P., & Boulanger, F. 1986, *A&A*, 170, 84
- Wakker, B. P., & Mathis, J. S. 2000, *ApJ*, 544, L107
- Wakker, B. P., & Savage, B. D. 2009, *ApJS*, 182, 378
- Wakker, B. P., & Schwarz, U. J. 1991, *A&A*, 250, 484
- Wakker, B. P., & van Woerden, H. 1991, *A&A*, 250, 509
- Wakker, B. P., & van Woerden, H. 1997, *ARA&A*, 35, 217
- Wakker, B. P., Howk, C., Schwarz, U. J., van Woerden, H., Beers, T., Wilhelm, R., Kalberla, P., & Danly, L. 1996, *ApJ*, 473, 834
- Wakker, B. P., Murphy, E.M., van Woerden, H., & Dame, T. M. 1997, *ApJ*, 488, 216
- Wakker, B. P., van Woerden, H., de Boer, K. S., & Kalberla, P. M. W. 1998, *ApJ*, 493, 762
- Wakker, B. P., Howk, J. C., Savage, B. D., van Woerden, H., Tufté, S. L., Schwarz, U. J., Benjamin, R., Reynolds, R. J., Peletier, R. F., & Kalberla, P. M. W. 1999, *Nature*, 402, 388
- Wakker, B. P., Kalberla, P. M. W., van Woerden, H., de Boer, K. S., & Putman, M. E. 2001, *ApJS*, 136, 537
- Wakker, B. P., Oosterloo, T. A., & Putman, M. E. 2002, *AJ*, 123, 1953
- Wakker, B. P., York, D. G., Howk, J. C., Barentine, J. C., Wilhelm, R., Peletier, R. F., van Woerden, H., Beers, T. C., Ivezić, Z., Richter, P. R., & Schwarz, U. J. 2007, *ApJ*, 670, L113
- Wakker, B. P., York, D. G., Wilhelm, R., Barentine, J. C., Richter, P., Beers, T. C., Ivezić, Z., & Howk, J. C. 2008, *ApJ*, 672, 298
- Wakker, B. P., Lockman, F. J., & Brown, J. 2011, *ApJ*, 728, 159
- Wakker, B. P., Savage, B. D., Fox, A. J., Benjamin, R., & Shapiro, P. 2012, *ApJ*, 749, 157
- Wannier, P., Wrixon, G. T., & Wilson, R. W. 1972, *A&A*, 18, 224
- Weiner, B. J., & Williams, T. B. 1996, *AJ*, 111, 1156
- Weiss, A., Heithausen, A., Herbstmeier, U., & Mebold, U. 1999, *A&A*, 344, 955
- Westmeier, T., Braun, R., & Thilker, D. A. 2005, *A&A*, 436, 101
- Westmeier, T., Brüns, C., & Kerp, J. 2008, *MNRAS*, 390, 1691
- White, S. D. M., & Rees, M. J. 1978, *MNRAS*, 183, 341
- Winkel, B., Kalberla, P. M. W., Kerp, J., & Flöer, L. 2010, *ApJS*, 188, 488
- York, D. G., et al. 2000, *AJ*, 120, 1579
- Zwaan, M. A. 2001, *MNRAS*, 325, 1142

13 Magnetic Fields in Galaxies

Rainer Beck · Richard Wielebinski

Max-Planck-Institut für Radioastronomie, Bonn, Germany

1	Introduction	643
2	Observational Methods	645
2.1	Optical and Far-Infrared Polarization	645
2.2	Synchrotron Emission	646
2.3	Magnetic Field Components	648
2.4	Faraday Rotation and Faraday Depolarization	650
2.5	Zeeman Effect	653
2.6	Field Origin and Amplification	653
3	Magnetic Fields in the Milky Way	657
3.1	Optical, Far-Infrared, and Sub-mm Polarization	657
3.2	Radio Continuum	658
3.2.1	All-Sky Surveys in Total Intensity	658
3.2.2	All-Sky Surveys in Linear Polarization	660
3.2.3	The Galactic Center	664
3.3	Faraday Rotation of Extragalactic Radio Sources and Pulsars	664
3.3.1	Extragalactic Radio Sources (EGRS)	665
3.3.2	Pulsars	667
3.4	Zeeman Effect	668
3.5	Modeling the Magnetic Field of the Milky Way	670
4	Galaxies	672
4.1	Optical Polarization, Infrared Polarization, and Zeeman Effect	672
4.2	Magnetic Field Strengths	674
4.3	The Radio–Infrared Correlation	678
4.4	Magnetic Field Structures in Spiral Galaxies	680
4.4.1	Ordered Fields	680
4.4.2	Regular Fields	684
4.5	Magnetic Fields in Barred Galaxies	690
4.6	Flocculent and Irregular Galaxies	693
4.7	Radio Halos	695
4.8	Interacting Galaxies	700
4.9	Galaxies with Jets	704
4.10	Elliptical and Dwarf Spheroidal Galaxies	706
5	Outlook	708

<i>Acknowledgments</i>	713
<i>Appendix</i>	713
A.1 Catalogue of Radio Polarization Observations of Nearby Galaxies	713
A.2 Links to the SKA Project and Its Precursor and Pathfinder Telescopes	718
<i>References</i>	718

Abstract: Most of the visible matter in the Universe is ionized so that cosmic magnetic fields are quite easy to generate and, due to the lack of magnetic monopoles, hard to destroy. Magnetic fields have been measured in or around practically all celestial objects, either by in situ measurements of spacecrafts or by the electromagnetic radiation of embedded cosmic rays, gas, or dust. The Earth, the Sun, solar planets, stars, pulsars, the Milky Way, nearby galaxies, more distant (radio) galaxies, quasars, and even intergalactic space in clusters of galaxies have significant magnetic fields, and even larger volumes of the Universe may be permeated by “dark” magnetic fields. Information on cosmic magnetic fields has increased enormously as the result of the rapid development of observational methods, especially in radio astronomy. In the Milky Way, a wealth of magnetic phenomena was discovered, which are only partly related to objects visible in other spectral ranges. The large-scale structure of the Milky Way’s magnetic field is still under debate. The available data for external galaxies can well be explained by field amplification and ordering via the dynamo mechanism. The measured field strengths and the similarity of field patterns and flow patterns of the diffuse ionized gas give strong indication that galactic magnetic fields are dynamically important. They may affect the formation of spiral arms, outflows, and the general evolution of galaxies. In spite of our increasing knowledge on magnetic fields, many important questions on the origin and evolution of magnetic fields, their first occurrence in young galaxies, or the existence of large-scale intergalactic fields remained unanswered. The present upgrades of existing instruments and several planned radio astronomy projects have defined cosmic magnetism as one of their key science projects.

Keywords: Cosmic rays, Dynamo action, Emission, Faraday rotation, Galactic Center, Galaxies: radio emission, Halos, Interstellar medium, Jets, Magnetic fields: origin, evolution, strength, structure, Milky Way: radio emission, Polarization, Pulsars, Radio telescopes, Spiral arms, Synchrotron, Zeeman effect

1 Introduction

The first report of a cosmic magnetic field outside the Earth was the result of a direct measurement of the Zeeman effect in the magnetic fields in sunspots of the Sun in 1908. In 1950, it was suggested that the observed cosmic rays would require magnetic fields for their creation and their containment within the Galaxy. Optical polarization observations were first successful in 1949. Polarization of optical and infrared emission can also be caused by elongated dust grains which are aligned in magnetic fields due to the Davis–Greenstein mechanism first described in 1951. This interpretation was not accepted for a long time in the optical astronomy community. With the advent of radio astronomy, this controversy could be resolved in favour of magnetic fields and an active study of magnetic fields could begin.

Radio astronomy began in 1932 with the detection of continuum radio emission from the Milky Way. It became quickly clear that the observed radio waves were of a nonthermal nature and an interpretation of this phenomenon was actively sought. This was given in 1949 – the radio emission is due to relativistic cosmic-ray electrons gyrating in magnetic fields, emitting radio waves by the synchrotron process – when the theory of synchrotron emission theory was developed. In particular, it was soon pointed out that synchrotron emission should be highly polarized. In fact, in homogenous magnetic fields, up to 75% linear polarization of the continuum emission can be expected. This suggestion was taken up by observers of optical radiation who found in 1954 that the Crab Nebula was highly polarized and hence emitting light through

the synchrotron process. The radio confirmation of the polarization of the Crab Nebula followed in 1957. The first definite detection of the linear polarization of the Galactic radio waves was published by in 1962. At the same time, the polarization of the bright radio galaxy Cygnus A and the Faraday rotation of the polarization angles of the linearly polarized radio emission in Centaurus A were detected. Observations at two frequencies of a section of the Milky Way showed that the interstellar medium of the Milky Way can also cause Faraday effect. During this exciting time of definite detections of interstellar and extragalactic magnetic fields by observations of linear polarization, the Zeeman effect of radio spectral lines proved to be more elusive. Several groups attempted to measure magnetic fields by this direct method. It was in 1968 that finally the Zeeman effect at radio wavelengths was successfully observed in the absorption profile of the HI line in the direction of Cassiopeia A. From this time onward, considerable data were collected on the distribution of magnetic fields in the Milky Way.

In the optical range, the polarization is produced by the different extinction along the minor and major axis of dust grains, while at far-infrared and submillimeter wavelengths, the elongated dust grains themselves emit polarized emission, which was first detected in the 1980s. Progress has been slow, until recently an increase in reliable data became possible with the advent of submillimeter telescopes on excellent sites and sensitive polarimeters.

The first suggestions about the presence of magnetic fields in nearby galaxies were made in 1958 based on observations of the polarization of stars in the Andromeda galaxy, M31. In 1967, observations of the linear polarization of diffuse starlight started in bright nearby galaxies. In 1970, the polarization of stars in the Magellanic Clouds implied the presence of magnetic fields in these neighboring galaxies. Low-frequency radio observations of galaxies showed nonthermal spectra and hence indicated the presence of magnetic fields. The first detection of the linear polarization of the radio emission from nearby galaxies in 1972 led the way to massive improvement on our knowledge of the morphology of magnetic fields in galaxies. These early radio observations were in good agreement with the early optical polarization studies of galaxies.

In this chapter, the status of our knowledge about the magnetic fields in our Milky Way and in nearby star-forming galaxies is summarized. Magnetic fields are a major agent in the interstellar and intra-cluster medium and affect the physical processes in various ways. They contribute significantly to the total pressure which balances the gas disk of a galaxy against gravitation. Magnetic reconnection is a possible heating source for the ISM and halo gas. They affect the dynamics of the turbulent interstellar medium (ISM) and the gas flows in spiral arms. The shock strength in spiral density waves is decreased and structure formation is reduced in the presence of a strong field. The interstellar fields are closely connected to gas clouds. Magnetic fields stabilize gas clouds and reduce the star-formation efficiency to the observed low values. On the other hand, magnetic fields are essential for the onset of star formation as they enable the removal of angular momentum from protostellar clouds via ambipolar diffusion. MHD turbulence distributes energy from supernova explosions within the ISM and drives field amplification and ordering via the dynamo mechanism. In galaxies with low star-formation activity or in the outer disks, the magnetorotational instability can generate turbulence and heat the gas. Magnetic fields control the density and distribution of cosmic rays in the ISM. Cosmic rays accelerated in supernova remnants can provide the pressure to drive a galactic outflow and generate buoyant loops of magnetic fields (called the Parker instability). Understanding the interaction between the gas and the magnetic field is a key to understand the physics of galaxy disks and halos and the evolution of galaxies.

The magnetic field of the Milky Way is of particular importance for experiments to detect *ultrahigh-energy cosmic rays* (UHECRs). Results from the first years of AUGER indicate that the

arrival directions of detected UHECRs with energies of more than 10^{19} eV show a statistically significant coincidence with the positions of known nearby active galaxies. This interpretation only holds if the deflections in the magnetic fields of the intergalactic medium and the Milky Way halo are not larger than a few degrees. However, little is known about the structure and strength of the magnetic field in the halo of our Milky Way and beyond.

There is one class of galaxies where magnetic fields play a crucial role: “active” galaxies which are governed by a central Black Hole. The formation of jets and radio lobes can only be understood with the presence of magnetic fields. The physics of these phenomena is quite different from that in “normal” star-forming galaxies and will not be discussed in this chapter.

Magnetic fields have also been detected in the intergalactic medium surrounding the galaxies in a cluster through observations of nonthermal diffuse radio halos and the Faraday effect of background radio sources seen through the cluster. These intracluster magnetic fields are probably generated by turbulent gas motions as the result of massive interactions between galaxies and the intracluster gas. Magnetic fields affect thermal conduction in galaxy clusters and hence their evolution. Outflows from galaxies may have magnetized the intergalactic medium so that the general intergalactic space may be pervaded with magnetic fields. Unfortunately, cosmic rays and dust grains are missing outside of galaxies and galaxy cluster, and magnetic fields remain invisible. Intracluster magnetic fields are also beyond the scope of this chapter.

Cosmological models of structure formation indicate that the intergalactic space is probably permeated by magnetic filaments. Galactic winds, jets from active galaxies, and interactions between galaxies can magnetize the intergalactic medium. The detection of magnetic fields in intergalactic filaments and observations of the interaction between galaxies and the intergalactic space is one of the important tasks for future radio telescopes. Until now, the arguments for the presence of magnetic fields in the distant Universe are based on observations of the nonthermal radio emission and Faraday rotation in galaxies at high redshift. Magnetic fields existed already in QSOs at epochs with redshifts of at least $z \approx 6$ and in starburst galaxies at redshifts of at least $z \approx 4$, but the earliest magnetic fields are yet to be discovered (▶ Sect. 5).

2 Observational Methods

As the methods of measuring of magnetic fields have been discussed widely in the literature, a short summary of the methods clarifies the present limitations.

2.1 Optical and Far-Infrared Polarization

Elongated, rotating dust grains can be aligned with their major axis perpendicular to the field lines by paramagnetic alignment (Davis and Greenstein 1951) or, more efficiently, by radiative torque alignment (Hoang and Lazarian 2008). When the particles are observed with their major axis perpendicular to the line of sight (and the field is oriented in the same plane), the different extinction along the major and the minor axis leads to polarization, with the E-vectors pointing parallel to the field. This is the basis to measure magnetic fields with optical and near-infrared polarization by observing individual stars or of diffuse starlight. Extinction is most efficient for grains of sizes similar to the wavelength. These small particles are aligned only in the medium between molecular clouds, not in the dense clouds themselves (Cho and Lazarian 2005).

The detailed physics of the alignment is complicated and depends on the magnetic properties of the particles. The degree of polarization p (in optical magnitudes) due to a volume element along the line of sight δL is given by Ellis and Axon 1978:

$$p = \frac{KB_{\perp}^2 \zeta \delta L}{N_{\text{H}} T_{\text{g}} T^{1/2}}$$

where T is the gas temperature

T_{g} is the grain temperature

N_{H} is the gas density

ζ is the space density of grains

B_{\perp} is the magnetic field strength perpendicular to the line of sight

Light can also be polarized by scattering, a process unrelated to magnetic fields. This contamination small when observing stars but needs to be subtracted from diffuse light, requiring multicolor measurements.

In the far-infrared (FIR) and submillimeter wavelength ranges, the emission of elongated dust grains is intrinsically polarized and scattered light is negligible. If the grains are again aligned perpendicular to the magnetic field lines, the E-vectors point perpendicular to the field. FIR polarimetry probes dust particles in the warm parts of molecular clouds, while sub-mm polarimetry probes grains with large sizes which are aligned also in the densest regions. The field strength can be crudely estimated from the velocity dispersion of the molecular gas along the line of sight and the dispersion of the polarization angles in the sky plane, the *Chandrasekhar–Fermi method* (Chandrasekhar and Fermi 1953), further developed for the case of a mixture of large-scale and turbulent fields by Hildebrand et al. (2009) and Houde et al. (2009).

2.2 Synchrotron Emission

Charged particles (mostly electrons) moving at relativistic speeds (cosmic rays) around magnetic fields lines on spiral trajectories generate electromagnetic waves. Cosmic rays in interstellar magnetic fields are the origin of the diffuse radio emission from the Milky Way (Fermi 1949; Kiepenheuer 1950). A cosmic-ray electron of energy E (in GeV) in a magnetic field with a component perpendicular to the line of sight of strength B_{\perp} (in μG) emits a smooth spectrum with a maximum at:

$$\nu_{\text{max}} \approx 4 \text{ MHz } E^2 B_{\perp}$$

where B_{\perp} is the strength of the magnetic field component perpendicular to the line of sight. For particles with a continuous power spectrum of electron energies, the maximum contribution at a given frequency comes from electrons with about twice lower energy so that ν_{max} becomes about $4\times$ larger.

The half-power lifetime of synchrotron-emitting cosmic-ray electrons is:

$$t_{\text{syn}} = 8.35 \cdot 10^9 \text{ year } B_{\perp}^{-2} E^{-1}$$

$$t_{\text{syn}} = 1.06 \cdot 10^9 \text{ year } B_{\perp}^{-1.5} \nu^{-0.5}$$

where B_{\perp} is measured in μG , E in GeV, and ν in GHz.

The emissivity σ from cosmic-ray electrons with a power-law energy spectrum in a volume with a magnetic field strength B_{\perp} is given by:

$$\sigma \sim N_0 \nu^{(\gamma+1)/2} B_{\perp}^{(1-\gamma)/2}$$

where ν is the frequency

N_0 is the density of cosmic-ray electrons per energy interval

γ is the spectral index of the power-law energy spectrum of the cosmic-ray electrons ($\gamma \approx -2.8$ for typical spectra in the interstellar medium of galaxies)

A source of size L along the pathlength has the intensity:

$$I_{\nu} \sim N_0 B_{\perp}^{(1-\gamma)/2} L$$

A power-law energy spectrum of the cosmic-ray electrons with the spectral index γ leads to a power-law synchrotron spectrum $I \sim \nu^{\alpha}$ with the spectral index $\alpha = (\gamma + 1)/2$. The initial spectrum of young particles injected by supernova remnants with $\gamma_0 \approx -2.2$ leads to an initial synchrotron spectrum with $\alpha_0 \approx -0.6$. These particles are released into the interstellar medium. A stationary energy spectrum with continuous injection and dominating synchrotron loss has $\gamma \approx -3.2$ and $\alpha \approx -1.1$. If the cosmic-ray electrons escape from the galaxy faster than within the synchrotron loss time, the stationary spectrum has $\gamma \approx (\gamma_0 - \delta) \approx -2.8$ and $\alpha \approx (\alpha_0 - \delta/2) \approx -0.9$, where δ is the exponent of the energy dependence of the electron diffusion coefficient ($D = D_0 (E/E_0)^{\delta}$, typically $\delta \approx 0.6$).

The energy densities of cosmic rays (mostly relativistic protons + electrons), of magnetic fields, and of turbulent gas motions, averaged over a large volume of the interstellar medium and averaged over time, are comparable (*energy equipartition*):

$$W_{\text{cr}} \sim \frac{B^2}{8\pi} \sim \frac{\rho V^2}{2}$$

where W_{cr} is the energy density of cosmic rays

$B^2/8\pi$ is the energy density of the total magnetic field

$\rho V^2/2$ is the energy density of turbulent gas motions with density ρ and velocity dispersion V

On spatial scales smaller than the diffusion length of cosmic-ray electrons (typically a few kpc) and on time scales smaller than the acceleration time of cosmic rays (typically a few million years), energy equipartition is not valid.

Equipartition between cosmic rays and magnetic fields allows us to estimate the total magnetic field strength:

$$B_{\text{eq}} \sim ((k + 1)I_{\nu}/L)^{2/(5-\gamma)}$$

This revised formula by Beck and Krause (2005) is based on integrating the energy spectrum of the cosmic-ray protons and assuming a ratio k between the number densities of protons and electrons in the relevant energy range. The revised formula may lead to significantly different field strengths than the classical textbook formula which is based on integration over the radio frequency spectrum. Note that the exponent of two seventh given in the minimum-energy formula in many textbooks is valid only for $\gamma = -2$. The widely used *minimum-energy* estimate of the field strength is smaller than B_{eq} by the factor $((1 - \gamma)/4)^{2/(5-\gamma)}$, hence similar to B_{eq} for $\gamma \approx -3$.

The above formula is valid for steep spectra with $\gamma < -2$. For flatter spectra, the integration over the energy spectrum of the cosmic rays diverges and the calculation of W_{cr} has to be restricted to a limited energy interval.

For electromagnetic particle acceleration mechanisms, the proton/electron density ratio k for GeV particles is $\approx 40\text{--}100$, which directly follows from their different masses. (For an electron-positron plasma, $k = 0$.) If energy losses of the electrons are significant, e.g., in strong magnetic fields or far away from their places of origin, k can be much larger, and the equipartition value is a lower limit of the true field strength. On the other hand, the nonlinear relation between I_ν and B_\perp may lead to an overestimate of the true field strength when using the equipartition estimate if strong fluctuations in B_\perp occur within the observed volume. Another uncertainty occurs if only a small volume of the galaxies is filled with magnetic fields. Nevertheless, the equipartition assumption provides a reasonable first-order estimate. Due to the small exponent in the formula, the dependence on the input parameters is weak so that even large uncertainties do not affect the result much. The magnetic energy density based on the equipartition estimate agrees well with that of the turbulent gas motions. Furthermore, estimates of the synchrotron loss time based on the equipartition assumption can well explain the extent of radio halos around galaxies seen edge-on (see [Sect. 4.6](#)). Finally, independent measurements of the field strength by the Faraday effect in “magnetic arms” leads to similar values ([Sect. 4.2](#)).

In our Galaxy, the accuracy of the equipartition assumption can be tested directly because there is independent information about the energy density and spectrum of local cosmic rays from in situ measurements and from γ -ray data, which are emitted by the protons via bremsstrahlung. Combination with the radio synchrotron data yields a local strength of the total field of $\approx 6 \mu\text{G}$ and $\approx 10 \mu\text{G}$ in the inner Galaxy. These values are similar to those derived from energy equipartition. A more precise estimate of field strengths requires will be possible from forthcoming γ -ray data.

Linear polarization is a distinct signature of synchrotron emission. The emission from a single electron gyrating in magnetic fields is elliptically polarized. An ensemble of electrons shows only very low circular polarization but strong linear polarization with the plane of the E vector normal to the magnetic field direction. The intrinsic degree of linear polarization p is given by:

$$p_0 = \frac{1 - \gamma}{7/3 - \gamma}$$

Considering galactic radio emission with $\gamma \approx -2.8$, a maximum of $p_0 = 74\%$ linear polarization is expected. In normal observing situations, the percentage polarization is reduced due to fluctuations of the magnetic field orientation within the volume traced by the telescope beam ([Sect. 2.3](#)) or by Faraday depolarization ([Sect. 2.4](#)). The observed degree of polarization is also smaller due to the contribution of unpolarized thermal emission which may dominate in star-forming regions.

2.3 Magnetic Field Components

The intensity of synchrotron emission is a measure of the number density of cosmic-ray electrons in the relevant energy range and of the strength of the *total magnetic field* component in the sky plane. Polarized emission emerges from *ordered fields*. As polarization vectors are ambiguous by 180° , they cannot distinguish *regular fields* with a constant direction within the telescope beam from *anisotropic fields* (generated from isotropic turbulent magnetic fields by

■ Table 13-1

Field components are their observational signatures

Field component	Notation	Property	Observational signature
Total field	B	3D	Total synchrotron intensity, corrected for inclination
Total field in sky plane	B_{\perp}	2D	Total synchrotron intensity
Turbulent field in sky plane	$B_{\text{turb},\perp}$	2D	Unpolarized synchrotron intensity, beam depolarization, Faraday depolarization
Turbulent field along line of sight	$B_{\text{turb},\parallel}$	1D	Faraday depolarization
Ordered field perpendicular to the line of sight	$B_{\text{ord},\perp}^2 = B_{\text{an},\perp}^2 + B_{\text{reg},\perp}^2$	2D	Polarized synchrotron intensity, optical polarization
Anisotropic field perpendicular to the line of sight	$B_{\text{an},\perp}$	2D	Polarized synchrotron intensity, optical polarization
Regular field perpendicular to the line of sight	$B_{\text{reg},\perp}$	2D	Polarized synchrotron intensity, optical polarization
Regular field along line of sight	$B_{\text{reg},\parallel}$	1D	Faraday rotation and depolarization, Zeeman effect

Note that anisotropic fields and regular fields perpendicular to the line of sight cannot be distinguished

compression or shear of gas flows) which have a preferred orientation, but frequently reverse their direction on small scales. Unpolarized synchrotron emission indicates isotropic *turbulent fields* with random directions which have been amplified and tangled by turbulent gas flows (► Table 13-1).

Magnetic fields in galaxies preserve their direction only over the coherence scale, which can be determined by field tangling or by turbulence. If N is the number of cells with the size of the turbulence scale within the volume observed by the telescope beam and if the coherence length is constant, wavelength-independent depolarization occurs (Burn 1966):

$$DP = \frac{P}{P_0} = N^{-1/2}$$

If the medium is pervaded by a turbulent field B_{turb} (unresolved field with randomly changing direction) plus an ordered field B_{ord} (regular and/or anisotropic) which has a constant orientation in the volume observed by the telescope beam, it follows for constant density of cosmic-ray electrons:

$$DP = \frac{1}{(1 + q^2)}$$

and for the equipartition case (Sokoloff et al. 1998):

$$DP = \frac{(1 + 3.5q^2)}{(1 + 4.5q^2 + 2.5q^4)}$$

where $q = B_{\text{turb},\perp}/B_{\text{ord},\perp}$ (components in the sky plane). This gives larger DP values (i.e., less depolarization) than for the former case.

2.4 Faraday Rotation and Faraday Depolarization

The linearly polarized radio wave is rotated by the Faraday effect in the passage through a magneto-ionic medium (see [Fig. 13-1](#)). This effect gives us another method of studying magnetic fields – their regular component along the line of sight. The rotation angle Φ induced in a polarized radio wave is given by:

$$\Phi = k\lambda^2 \int n_e B_{\parallel} dl$$

with λ wavelength of observation

n_e thermal electron density

B_{\parallel} strength of the regular magnetic field component along the line of sight

dl pathlength along the magnetic field

and k is a constant (see below)

In practice, the parameter Faraday Depth (FD) is used (Burn 1966):

$$\Phi = FD\lambda^2 \text{ where } FD = 0.81 \int n_e B_{\parallel} dl > L(\text{rad m}^{-2}) \quad (13.1)$$

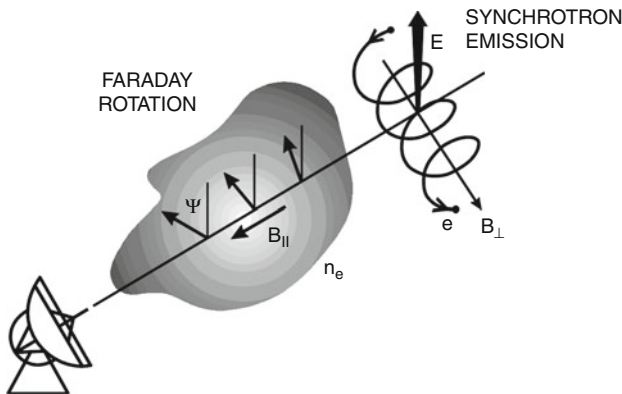
with n_e thermal electron density in cm^{-3}

B_{\parallel} regular field strength in μG

L pathlength in parsec

Note that the observable quantity Rotation Measure ($RM = \Delta\Phi/\Delta\lambda^2$) is identical to the physical quantity FD only in the rare cases when Φ is a linear function of λ^2 . If the rotating region is located in front of the emitting region (“Faraday screen”), $RM = FD$. In case of a single emitting and rotating region, $RM \approx FD/2$ if Faraday depolarization (see below) is small.

As Faraday rotation angle is sensitive to the sign of the field direction, only regular fields give rise to Faraday rotation, while anisotropic and turbulent fields do not. For typical plasma densities and regular field strengths in the interstellar medium of galaxies, Faraday rotation becomes significant at wavelengths larger than a few centimeters. Only in the central regions of galaxies,



■ Fig. 13-1

Synchrotron emission and Faraday rotation

Faraday rotation is strong already at 1–3 cm wavelengths. Measurements of the Faraday rotation angle from multiwavelength observations allow determination the strength and direction of the regular field component along the line of sight. Its combination with the total intensity and the polarization pseudovectors yields in principle the three-dimensional picture of galactic magnetic fields and the three field components – regular, anisotropic, and turbulent.

By definition, the regular magnetic fields point toward the observer when $RM > 0$. The quantity $\langle n_e B_{\parallel} \rangle$ is the average of the product $(n_e B_{\parallel})$ along the line of sight which generally is not equal to the product of the averages $\langle n_e \rangle \langle B_{\parallel} \rangle$ if fluctuations in n_e and B_{\parallel} are correlated or anticorrelated. As a consequence, the field strength $\langle B_{\parallel} \rangle$ cannot be easily determined from RM even if additional information about $\langle n_e \rangle$ is available, e.g., from pulsar dispersion measures (► Sect. 3.3).

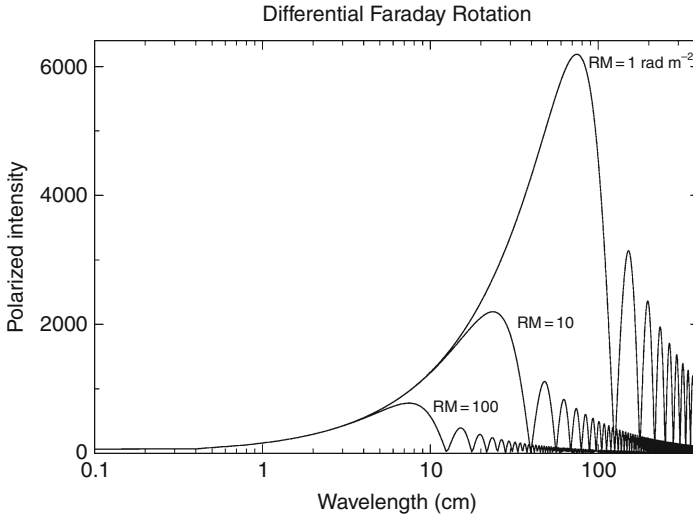
Measurement of RM needs polarization observations in at least three frequency channels with a large frequency separation in a frequency range where Faraday depolarization is still small. In case of strong Faraday depolarization (see below), the polarization angle Φ is no longer a linear function of λ^2 . Large deviations from the λ^2 law can also occur if several emitting and Faraday-rotating sources are located within the volume traced by the telescope beam. In such cases, the local RM measured over a small wavelength range strongly fluctuates with wavelength, and polarization data cubes with many frequency channels (spectro-polarimetry) are needed to allow application of *RM Synthesis* (Brentjens and de Bruyn 2005).

RM Synthesis Fourier-transforms the complex polarization data from a limited part of the λ^2 space into a data cube with FD as the third coordinate, similar to synthesis of images from interferometric telescopes. The total span and the distribution of the frequency channels and the channel width define the *Rotation Measure Spread Function (RMSF)*, which allows cleaning of the RM data cube, similar to cleaning of synthesis data from interferometric telescopes (Heald 2009). RM Synthesis is able to separate FD components from distinct foreground and background regions and hence in principle to measure the 3D structure of the magnetized medium. If the structure is relatively simple, i.e., a few emitting and Faraday-rotating regions, “Faraday tomography” is possible.

In a region containing cosmic-ray electrons, thermal electrons, and purely regular magnetic fields, wavelength-dependent Faraday depolarization occurs because the polarization planes of waves from the far side of the emitting layer are more rotated than those from the near side. This effect is called *differential Faraday rotation* and is described (for one single layer with a symmetric distribution of thermal electron density and field strength along the line of sight) by (Burn 1966):

$$DP = \frac{|\sin(2RM\lambda^2)|}{|(2RM\lambda^2)|}$$

where RM is the observed rotation measure, which is half of the total rotation measure through the whole layer. DP varies periodically with wavelength. With $|RM| = 100 \text{ rad m}^{-2}$, typical for normal galaxies, DP has zero points at wavelengths of $(12.5 \sqrt{n}) \text{ cm}$, where $n = 1, 2, \dots$ (► Fig. 13-2). At each zero point, the polarization angle jumps by 90° . Observing at a fixed wavelength hits zero points at certain values of the intrinsic RM, giving rise to *depolarization canals* along the level lines of RM. At wavelengths just below that of the first zero point in DP, only the central layer of the emitting region is observed because the emission from the far side and that from the near side cancel (their rotation angles differ by 90°). Beyond the first zero point, only a small layer on the near side of the disk remains visible. Applying RM Synthesis to multichannel observations yields a tomographic picture through the region.



■ Fig. 13-2

Wavelength of maximum polarized emission for a synchrotron spectrum with spectral index $\alpha = -0.9$ and depolarization by differential Faraday rotation at the level of $|RM|$ (Arshakian and Beck 2011)

Turbulent fields also cause wavelength-dependent depolarization, called *Faraday dispersion* (Sokoloff et al. 1998). For an emitting and Faraday-rotating region (internal dispersion):

$$DP = \frac{(1 - \exp(-S))}{S}$$

where $S = 2\sigma_{RM}^2\lambda^4$.

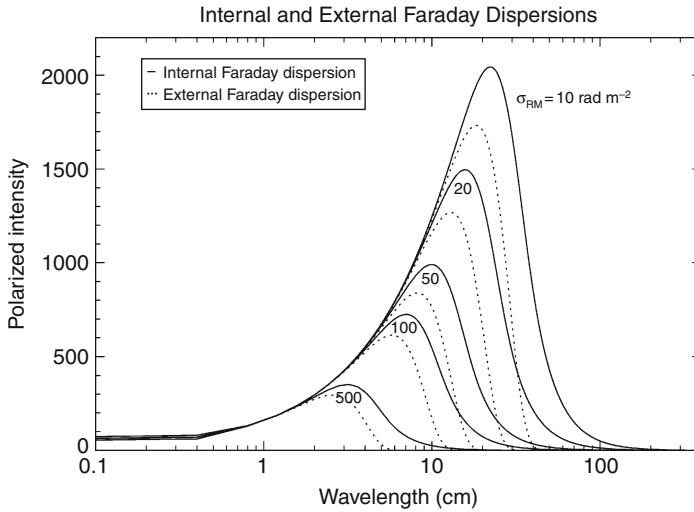
σ_{RM}^2 is the dispersion in rotation measure and depends on the turbulent field strength along the line of sight, the turbulence scale, the thermal electron density, and the pathlength through the medium. The main effect of Faraday dispersion is that the interstellar medium becomes “optically thick” for polarized radio emission beyond a wavelength, depending on σ_{RM} (► Fig. 13-3), and only a front layer remains visible in polarized intensity. Galaxy halos and intracluster media have typical values of $\sigma_{RM} = 1 - 10 \text{ rad m}^{-2}$, while galaxy disks have $\sigma_{RM} = 10 - 100 \text{ rad m}^{-2}$. Centers of galaxies can have even higher dispersions. ► Figure 13-3 shows the optimum wavelength ranges to detect polarized emission for these regions.

Regular fields in a non-emitting foreground *Faraday screen* do not depolarize, while turbulent fields do (external Faraday dispersion). For extended sources:

$$DP = \exp(-S)$$

Unresolved *RM gradients* within the beam also lead to depolarization, similar to Faraday dispersion.

Faraday depolarization can also be classified as *depth depolarization* (differential Faraday rotation, Faraday dispersion along the line of sight) and *beam depolarization* (RM gradients, Faraday dispersion in the sky plane). Both types occur in emitting regions, while in non-emitting Faraday screens only beam depolarization occurs.



■ Fig. 13-3

Wavelength of maximum polarized emission for a synchrotron spectrum with spectral index $\alpha = -0.9$ and depolarized by Faraday dispersion at the level of σ_{RM} . *Solid curve*: internal Faraday dispersion within an emitting source; *dotted curve*: external Faraday dispersion in a foreground object (Arshakian and Beck 2011)

2.5 Zeeman Effect

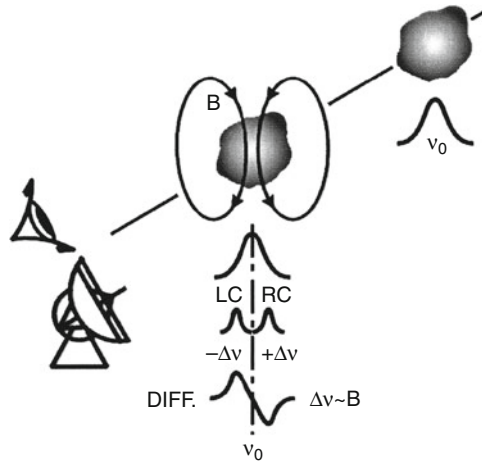
The Zeeman effect is the most direct method of remote sensing of magnetic fields. It has been used in optical astronomy since the first detection of magnetic fields in sunspots of the Sun. The radio detection was first made in the HI line. In the presence of a magnetic field B_{\parallel} along the line of sight, the line at the frequency ν_0 is split into two components (*longitudinal Zeeman effect*) (see ● Fig. 13-4):

$$\nu_0 \pm \frac{e B_{\parallel}}{4\pi m c}$$

where e , m , and c are the usual physical constants. The two components are circularly polarized of the opposite sign. The frequency shift is minute, e.g., 2.8 MHz G^{-1} for the HI line. More recent observation of the OH or H_2O lines used the higher frequency shifts of these molecular line species. In magnetic fields perpendicular to the line of sight, two shifted lines together with the main unshifted line, all linearly polarized. This *transversal Zeeman effect* is much more difficult to observe and has not yet been detected in the interstellar medium.

2.6 Field Origin and Amplification

The origin of the first magnetic fields in the Universe is still a mystery (Widrow 2002). The generation of the very first “seed” fields needs a continuous separation of electric charges, e.g., by the *Biermann battery* or the *Weibel instability* (Lazar et al. 2009). The Biermann battery



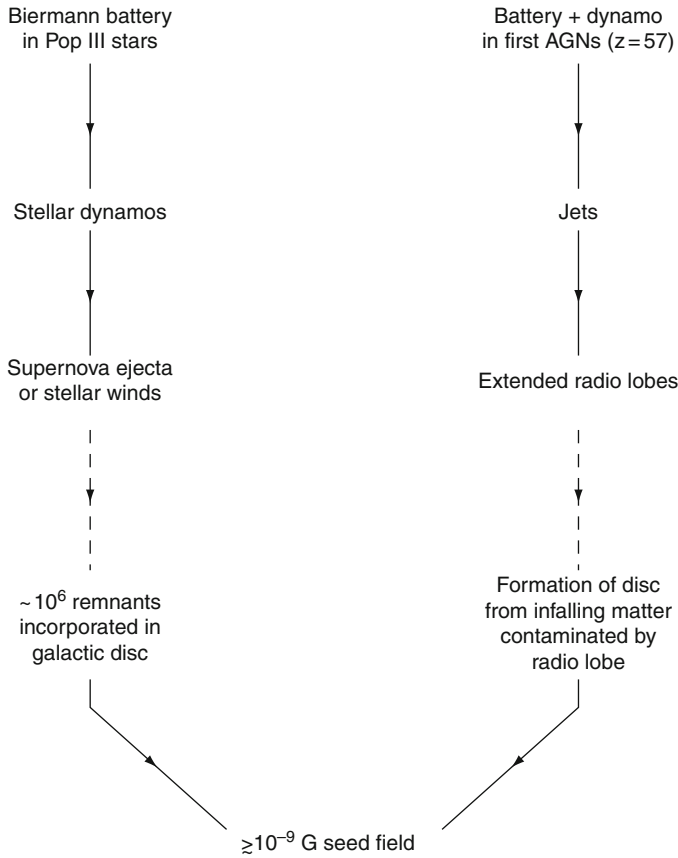
■ Fig. 13-4

The longitudinal Zeeman effect, splitting of a line into two components with opposite circular polarization

may generate a field of $\leq 10^{-20}$ G in the first galaxies or stars. A large-scale intergalactic field of $\leq 10^{-12}$ G may be generated in the early Universe and could also serve as a seed field in protogalaxies. This is consistent with the average strength of intergalactic fields of $\geq 10^{-16}$ G, derived from high-energy γ -ray observations with HESS and FERMI, by assuming that the secondary particles are deflected by the intergalactic fields (Neronov and Vovk 2010). However, such a field is hard to maintain because the galaxy rotates differentially so that field lines get strongly wound up, in contrast to the observations (Shukurov 2005). Moreover, a coherent large-scale field as observed, e.g., in M31 cannot be explained by the primordial field model. The same is true for kinematical models of field generation by induction in shearing and compressing gas flows, which generate fields with a few kpc coherence length and frequent reversals.

More promising is the magnetization of protogalaxies to $\geq 10^{-9}$ G by field ejection from the first stars or the first black holes (► Fig. 13-5), followed by dynamo action. The dynamo transfers mechanical into magnetic energy. It amplifies and /or orders a seed field. The *small-scale* or *fluctuation dynamo* does not need general rotation, only turbulent gas motions (e.g., Brandenburg and Subramanian 2005). The source of turbulence can be thermal virialization in protogalactic halos or supernovae in the disk or the *magnetorotational instability (MRI)* (e.g., Rüdiger and Hollerbach 2004). Within less than 10^9 years, even weak seed fields are amplified to the energy density level of turbulence and reach strengths of a few μG .

The *mean-field* or $\alpha - \Omega$ dynamo is driven by turbulent gas motions from supernova explosions or cosmic-ray driven Parker loops (α) and by differential rotation (Ω), plus magnetic diffusivity (η) (e.g., Beck et al. 1996; Parker 1979; Ruzmaikin et al. 1988). It generates a large-scale (“mean”) regular field from the turbulent field in a typical spiral galaxy within a few 10^9 years. If the small-scale dynamo already amplified turbulent fields of a few μG in the protogalaxy, the mean-field dynamo is needed only for the organization of the field (“order out of chaos”). The field pattern is described by modes of different azimuthal symmetry in the disk and vertical symmetry or antisymmetry perpendicular to the disk plane. Several modes can

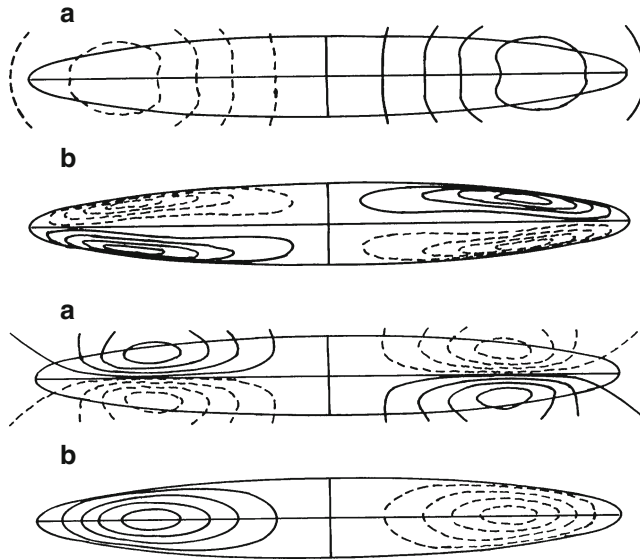


■ Fig. 13-5

Origin of seed fields in protogalaxies (Rees 2005). The last stage can be replaced by the small-scale dynamo

be excited in the same object. In almost spherical, rotating bodies like stars, planets, or galaxy halos, the strongest mode consists of a toroidal field component with a sign reversal across the equatorial plane (vertically antisymmetric or “odd” parity mode A0) and a dipolar poloidal field component with field lines crossing the equatorial plane. The halo mode can also be oscillatory and reverse its parity with time (e.g., causing the cycle of solar activity). The oscillation timescales (if any) are very long for galaxies and cannot be determined by observations.

In flat, rotating objects like galaxy disks, the strongest mode consists of a toroidal field component, which is symmetric with respect to the equatorial plane and has the azimuthal symmetry of an axisymmetric spiral in the plane without sign reversals (vertically symmetric or “even” parity mode S0), and a weaker poloidal field component of quadrupolar symmetry with a reversal of the vertical field component across the equatorial plane (▶ Fig. 13-6). The next higher azimuthal mode is of bisymmetric spiral shape (even mode S1) with two sign reversals in the plane, followed by more complicated modes. The pitch angle of the spiral field is determined by the rotation curve of the galaxy, the turbulent velocity and the scale height of the



■ Fig. 13-6

Poloidal field lines (a) and contours of constant toroidal field strength (b) for the simplest version of a dipolar (top) and quadrupolar (bottom) dynamo field (Stix 1975). More realistic dynamo fields can have many “poles” (reproduced with permission © ESO)

warm diffuse gas (Shukurov 2005). The field in fast-rotating galaxies has a small pitch angle of about 10° , while slow differential rotation or strong turbulence leads to a larger pitch angle of $20\text{--}30^\circ$.

In principle, the halo and the disk of a galaxy may drive different dynamos and host different field modes. However, there is a tendency of “mode slaving,” especially in case of outflows from the disk into the halo. The more dynamo-active region determines the global symmetry so that the halo and disk field should have the same parity (Moss et al. 2010). This is confirmed in external galaxies (▶ Sect. 4.7), while our Milky Way seems to be different (▶ Sect. 3.5).

The ordering time scale of the mean-field dynamo depends on the size of the galaxy (Arshakian et al. 2009). Large galaxies did not yet have sufficient time to build up a fully coherent regular field and may still host complicated field patterns, as often observed. The field ordering may also be interrupted by tidal interactions or merging with another galaxy, which may destroy the regular field and significantly delays the development of coherent fields (▶ Sect. 5). Strong star formation as the result of a merger event or mass inflow amplifies the turbulent field and can suppress the mean-field dynamo in a galaxy if the total star-formation rate is larger than about 20 solar masses per year.

The mean-field dynamo generates large-scale helicity with a nonzero mean in each hemisphere. As total helicity is a conserved quantity, the dynamo is suppressed by the small-scale fields with opposite helicity, unless these are removed from the system (e.g., Vishniac et al. 2003). Hence, outflow with a moderate velocity or diffusion is essential for an effective mean-field dynamo. This effect may relate the efficiency of dynamo action to the star-formation rate in the galaxy disk (▶ Sect. 4.6). Mean-field dynamo models including outflows with moderate

velocities can also generate X-shaped fields (Moss et al. 2010). For fast outflows, the advection time for the field becomes smaller than the dynamo amplification time so that the dynamo action is no longer efficient.

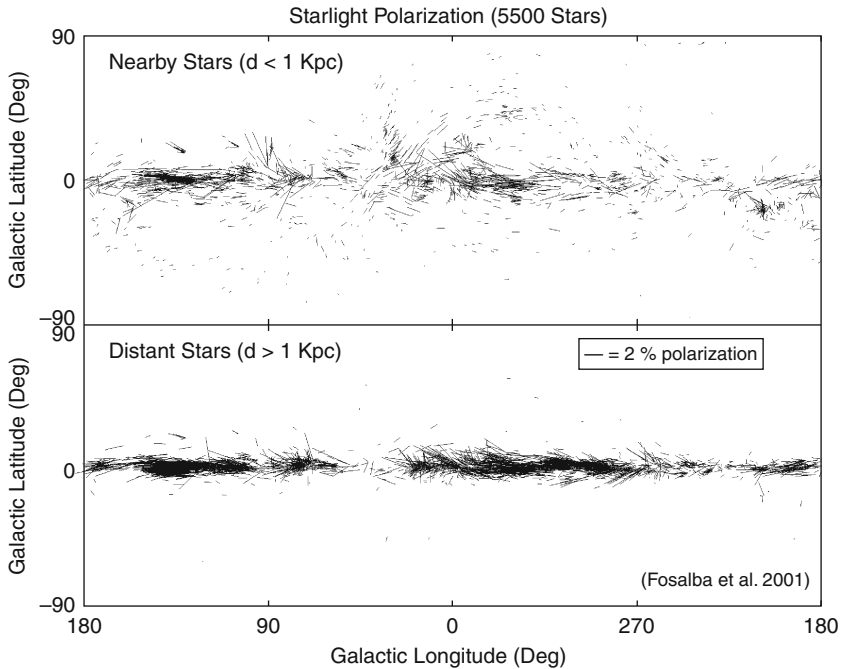
There are several more unsolved problems with dynamo theory (Vishniac et al. 2003). The “mean-field” model is simplified because it assumes a dynamical separation between the small and the large scales. The large-scale field is assumed to be smoothed by turbulent diffusion, which requires fast and efficient field reconnection. One of the main future tasks is to compute the “mean” quantities α and η from the small-scale properties of the interstellar medium, which is only possible with numerical modeling. MHD simulations with 100 pc resolution of a dynamo driven by supernovae (Gressel et al. 2008) or by the buoyancy of cosmic rays (Hanasz et al. 2009) confirm the overall description of the mean-field model. Improved models need a spatial resolution of smaller than the turbulence scale, hence ≈ 10 pc, must include the whole rotating galaxy, include the halo and consider all relevant physical effects. The multiphase interstellar medium has also to be taken into account (de Avillez and Breitschwerdt 2005). Rapid progress in modeling galactic magnetic fields can be expected in near future.

While the predictions of the dynamo model have been generally confirmed by present-day observations (► Sects. 4.3 and ► 4.6), the primordial model of field amplification is less developed than the dynamo model and is not supported by the data. A wound-up large-scale seed field can generate only the even bisymmetric mode (S1) or the odd dipolar mode (A0), both of which were not observed so far. On the other hand, the number of galaxies with a well-determined field structure is still limited (Appendix). Future radio telescopes will be able to decide whether the dynamo or the primordial model is valid or whether a new model has to be developed.

3 Magnetic Fields in the Milky Way

3.1 Optical, Far-Infrared, and Sub-mm Polarization

The earliest optical polarization observations in 1949 were interpreted to be due to dust alignment in magnetic fields and hence a tracer of magnetic fields in galaxies. It took some time to convince the optical community that the polarization was due to dust grains aligned in magnetic fields. The radio polarization observations (► Sect. 3.2) confirmed the magnetic explanation. A large catalogue of the polarization of stars was made by Behr (1961). This work continued in the southern skies, as well as other observers, culminating in an all-sky catalogue of Mathewson and Ford (1970a) with 1,800 entries and Axon and Ellis (1976) with 5,070 entries. The general conclusion of this work, that there is a magnetic field aligned along the Galactic plane, still holds today. A very homogeneous region of alignment, with high polarization values, was seen toward the anticenter (Galactic longitude $l \approx 140^\circ$). Well-aligned magnetic field vectors are also seen along the North Polar Spur that extends in to the northern halo from $l \approx 30^\circ$. These early observations were possible for nearby stars, a few at a maximal distance of 4 kpc. A more recent compilation of 9,286 stars, collected by Heiles (2000) and discussed by Fosalba et al. (2002) (► Fig. 13-7), included some stars out to ≈ 8 kpc. In view of these distance limitations, it is not possible on the basis of optical polarization alone to model the magnetic field of the Milky Way.



■ Fig. 13-7

Optical starlight polarization in the Galactic plane for two distance intervals (Fosalba et al. 2002)

Polarization observations of the diffuse far-infrared or sub-mm emission in the Milky Way are restricted to dense molecular/dust clouds. The Chandrasekhar–Fermi method (● Sect. 2.1) gives field strengths of a few mG, similar to Zeeman measurements of OH maser lines in other dense clouds (● Sect. 3.4). Interferometric observations in the sub-mm range with sub-parsec resolution reveals hourglass morphologies in the envelopes of the dust cores of ultra-compact HII regions (Tang et al. 2009). The supercritical cores seem to collapse in a subcritical envelope supported by strong magnetic fields, suggesting that ambipolar diffusion plays a key role in the evolution of the cloud. The correlation of the field orientation in the intercloud medium on a scale of several 100 pc, derived from optical polarization, with that in the cloud core on a scale of less than 1 pc, derived from sub-mm polarimetry, further indicates that the fields are strong and preserve their orientation during cloud formation (Li et al. 2009).

3.2 Radio Continuum

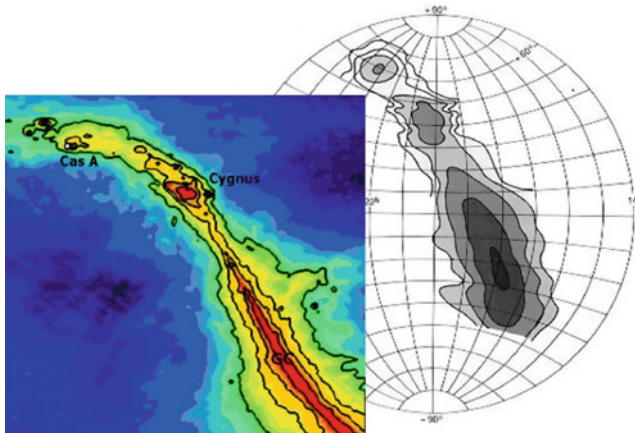
3.2.1 All-Sky Surveys in Total Intensity

The radio continuum emission of the Milky Way and star-forming galaxies at frequencies below 10 GHz mostly originates from the synchrotron process and hence traces the distribution of magnetic fields and cosmic rays. The contribution of thermal radio emission is generally small, except in bright star-forming regions. Only at frequencies higher than 10 GHz, the

thermal emission may dominate locally. At frequencies below about 300 MHz, absorption of synchrotron emission by thermal gas can become strong. Hence, the observation of total radio continuum intensity in the frequency range of about 300 MHz–10 GHz is a perfect method to investigate magnetic fields. Since the observed intensity is the integral from many emission areas along the line of sight, its interpretation is not always simple. Furthermore, the angular resolution of all-sky surveys is limited and hence cannot show the details of extended sources (► Fig. 13-10).

Numerous radio continuum surveys were made in the early days of radio astronomy (► Table 13-2). The early all-sky surveys showed the Galactic emission with a maximum toward the Galactic center, the band of emission along the Galactic plane, maxima in the tangential directions of the local spiral arm: Cygnus ($l \approx 80^\circ$) in the northern and Vela ($l \approx 265^\circ$) in the southern skies and some “spurs” of emission. In addition, a few strong extragalactic sources were seen superposed on the Galactic emission.

The analysis of total synchrotron emission gives an equipartition strength of the total field of $6 \pm 2 \mu\text{G}$ in the local neighborhood and $10 \pm 3 \mu\text{G}$ at 3 kpc radius (Berkhuijsen, in Beck 2001).



■ Fig. 13-8

The early sky map at 160 MHz of Grote Reber (1944) (black-white) and a recent color 1.4 GHz map (courtesy Wolfgang Reich)

■ Table 13-2

All-sky or all-hemisphere radio total intensity surveys

Frequency	Beam	Reference
45 MHz	$\approx 4^\circ$	Guzmán et al. 2011
150 MHz	3.6°	Landecker and Wielebinski 1970
408 MHz	2°	Haslam et al. 1982
1.4 GHz	0.6°	Reich 1982; Reich and Reich 1986; Reich et al. 2001
2.3 GHz	0.33°	Jonas et al. 1998 (Southern hemisphere)
2.7 GHz	0.33°	Reif et al. 1987 (Northern hemisphere)
23–94 GHz	0.8° – 0.2°	Hinshaw et al. 2009

The radial scale length of the total field is about 12 kpc. These results are similar to those in external galaxies (▶ Sect. 4.2).

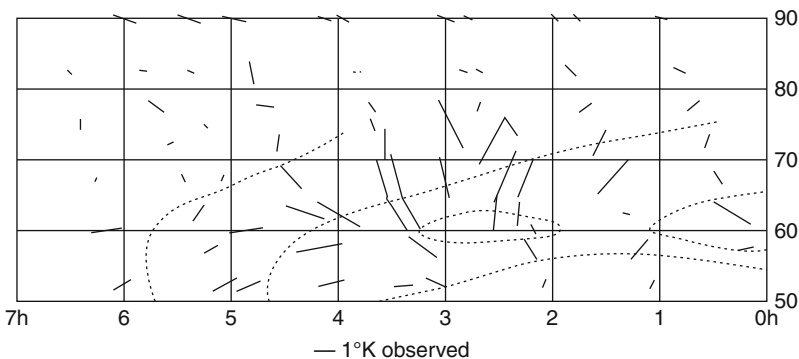
The angular resolution has improved so that, at present, all-sky surveys with resolution of under 1° are available. At 1.4 GHz, the surveys delineated many extended Galactic sources (HII regions, SNRs) seen along the Galactic plane. Some extragalactic sources like Centaurus A, Virgo A, Cygnus A, and the Magellanic Clouds are also clearly seen in the all-sky survey. Surveys at 45 MHz covered most of the sky with medium angular resolution. At these low frequencies, absorption of the synchrotron emission by ionized gas takes place near the Galactic plane.

The WMAP satellite surveys at frequencies from 23 to 94 GHz (Bennett et al. 2003; Hinshaw et al. 2009) gave us a new view of the radio continuum sky at high radio frequencies. At the highest WMAP frequencies, mainly thermal emission originating in interstellar dust is observed. An additional component due to spinning dust has been postulated (Draine and Lazarian 1998) to be seen in the 10–100 GHz frequency range. This spinning dust component has recently been confirmed (Dobler et al. 2009) in the WMAP data set.

There is a large gap between the lower frequency all-sky surveys and the high-frequency data. A 5 GHz all-sky survey with compatible angular resolution but also good sensitivity is badly needed. ▶ Table 13-2 lists the all-sky surveys with the best angular resolution at a given frequency.

3.2.2 All-Sky Surveys in Linear Polarization

Linear polarization of the continuum emission is a more direct indicator of magnetic fields because there is no confusing thermal component. However, linear polarization is subject to Faraday effects (▶ Sect. 3.3). After the first detections of polarized Galactic radio waves in 1962 (see ▶ Fig. 13-9), several all-sky polarization surveys were made (▶ Table 13-3). The early polarization surveys did not have sufficient angular resolution to elucidate many details. These surveys were made at the low radio frequency of 408 MHz where Faraday effects are considerable. A multifrequency collection of polarization data for the northern sky was published by Brouw and Spoelstra (1976), albeit not fully sampled.



■ Fig. 13-9

First detection of polarized synchrotron emission (E-vectors) in the Milky Way at 408 MHz (Wielebinski et al. 1962)

■ **Table 13-3**

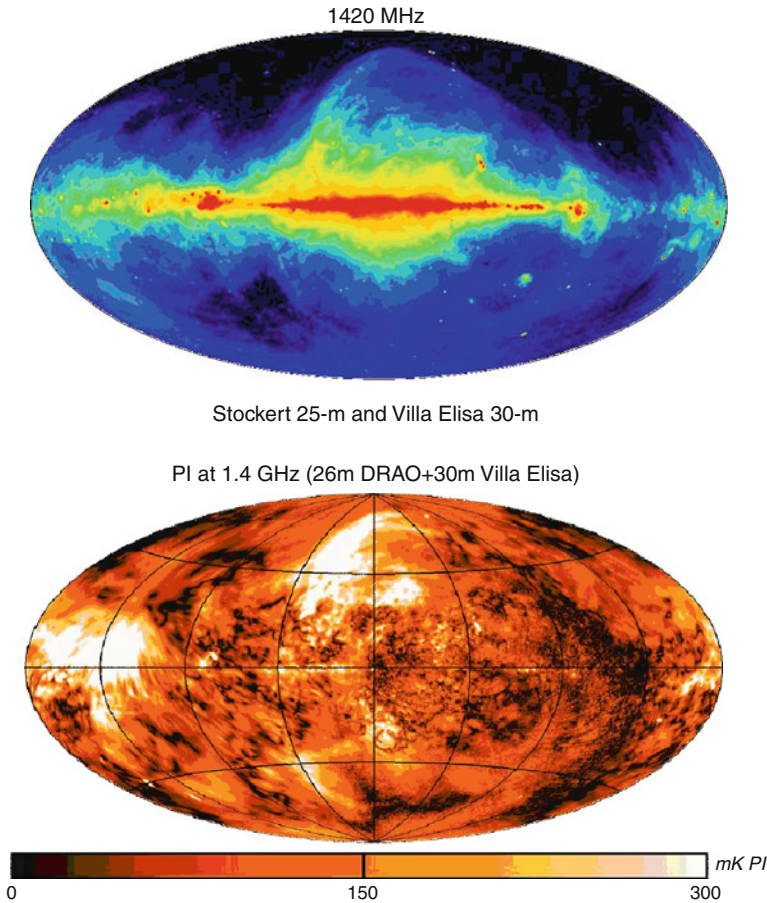
All-sky or all-hemisphere radio polarization surveys

Frequency	Beam	Reference
408 MHz	7.5°	Wielebinski et al. 1962 (Northern hemisphere)
	2°	Berkhuijsen and Brouw 1963 (Northern hemisphere)
	7.5°	Wielebinski and Shakeshaft 1964 (Northern hemisphere)
	≈ 1°	Mathewson and Milne 1965 (Southern hemisphere)
300 MHz–1.8 GHz	30′–60′	Wolleben et al. 2009, 2010a (Northern hemisphere)
1.4 GHz	36′	Testori et al. 2008; Wolleben et al. 2006
1.4 GHz	≈ 13′	Rudnick and Brown 2009 (Northern hemisphere)
23–94 GHz	0.8°–0.2°	Hinshaw et al. 2009; Kogut et al. 2007

Major progress was achieved by Wolleben et al. (2006) and Testori et al. (2008) who mapped the whole sky in linear polarization at 1.4 GHz with an angular resolution of 36 arcmin (► Fig. 13-10). Several polarization maxima are seen, e.g., toward the “Fan region” at $l \approx 140^\circ$, $b \approx 10^\circ$, where the line of sight is oriented perpendicular to the local spiral arm. The “North Polar Spur” (NPS) emerges from the Galactic plane at $l \approx 30^\circ$ as well as additional spur-like features are the results of magnetic fields compressed by expanding supernova remnants. In particular, the NPS can be followed, in polarization, to the southern sky. Toward the inner Galaxy (Galactic longitude $90^\circ > l > 270^\circ$, Galactic latitude $|b| < 30^\circ$), strong turbulence in the polarized intensity is seen due to Faraday effects on small scales (► Sect. 2.4). The NRAO VLA sky survey (NVSS) has also recently been analyzed in polarization (Rudnick and Brown 2009). All-sky polarization data at 23 GHz was published by the WMAP team (Hinshaw et al. 2009; Kogut et al. 2007). There is good agreement between the 23 GHz and the 1.4 GHz polarization maps in the polarization features away from the Galactic plane, but the high-frequency map shows less Faraday depolarization toward the inner Galaxy and near the plane. Another major survey has been started which will cover the whole sky at frequencies between 300 MHz and 1.8 GHz and will allow the measurement the RM of the diffuse emission over the whole sky (Wolleben et al. 2009). A summary of all-sky polarization surveys is given in Reich (2006) and in ► Table 13-3.

Galactic plane surveys have been made from the earliest days of radio astronomy to delineate the extended Galactic sources like supernova remnants and HII regions, usually with no linear polarization data (► Table 13-4). Many of the published Galactic plane surveys between 22 MHz and 10 GHz cover only a narrow strip along the Galactic plane in the inner Galaxy. Total intensity surveys at several frequencies were used to separate the thermal HII regions (with a flat radio spectrum) from the steep-spectrum nonthermal sources (supernova remnants). From the total intensity surveys, numerous previously unknown supernova remnants could be identified.

Since nonthermal sources are polarized, it was obviously necessary to map the Galactic plane also in linear polarization. The first step in the evolution of our knowledge about the polarization of the Galactic plane was the 2.7 GHz survey by Junkes et al. (1987), followed by the surveys of the southern Galactic plane at 2.3 GHz (Duncan et al. 1995) and the northern counterpart at 2.7 GHz (Duncan et al. 1999), which covered a relatively wide strip ($|b| < 5^\circ$) around the plane. Early high-resolution observations by Wieringa et al. (1993) showed that a lot of small-scale polarization is present in the Galactic emission which is unrelated to any structures in total intensity. The next major development is the Effelsberg Medium Latitude Survey (EMLS) at 1.4 GHz that will ultimately cover $\pm 20^\circ$ distance from the Galactic plane (Reich et al. 2004; Uyanıker et al. 1999). A section of the southern Galactic plane has been



■ Fig. 13-10

All-sky surveys in total intensity (*top*) and polarized intensity (*bottom*) at 1.4 GHz (Reich 1982; Testori et al. 2008; Wolleben et al. 2006)

mapped at 1.4 GHz with arcminute resolution (Gaensler et al. 2001; Haverkorn et al. 2006), complemented on the northern sky by the DRAO survey (Landecker et al. 2010; Taylor et al. 2003) (► Fig. 13-11). A survey of a $5^\circ \times 90^\circ$ strip along the Galactic meridian $l = 254^\circ$ with the Parkes telescope at 2.3 GHz is underway (Carretti et al. 2010).

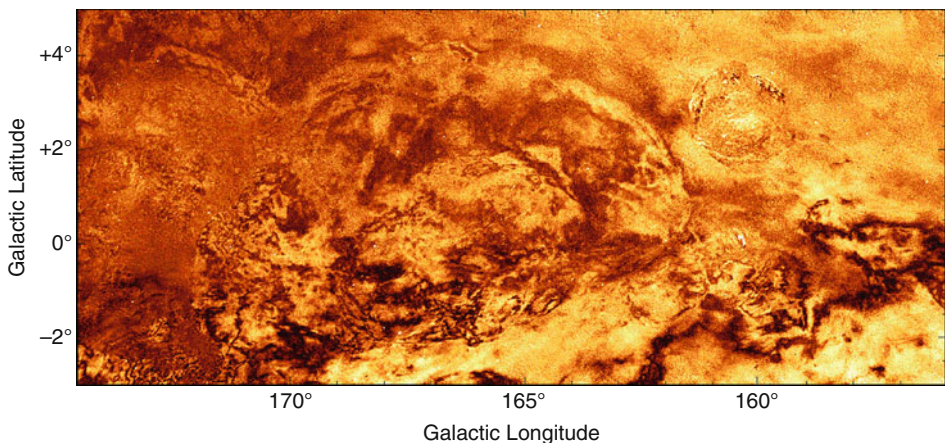
In all the above mentioned surveys *Faraday effects* (► Sect. 2.4) play an important role. At frequencies of 1.4 GHz and below, Faraday rotation generates small-scale structures in polarization which are not related to physical structures. Even at 5 GHz, Faraday rotation plays an important role near the Galactic plane ($|b| < 5^\circ$). With high enough angular resolution, Faraday rotation leads to complete depolarization at certain values of Faraday rotation measure (RM) (► Fig. 13-2), showing up as “canals” in the maps of polarized intensity (e.g., Haverkorn et al. 2003, 2004 and ► Fig. 13-11). However, a careful determination of the extended polarized background is necessary for a reliable determination of polarized intensity, polarization angles, and RM. When this “absolute calibration” is done, most of the “canals” disappear. As a second new phenomenon, “Faraday Screens” were discovered (e.g., Gray et al. 1998; Schnitzeler et al. 2009;

■ Table 13-4

Radio surveys of the Galactic plane with angular resolutions of a few arcminutes

Frequency	Beam	Area	Reference
325 MHz	$\approx 4'$	selected areas	Wieringa et al. 1993
350 MHz	$\approx 5'$	selected areas	Haverkorn et al. 2003, 2004
408 MHz	$\approx 1'$	$147.3^\circ > l > 74.2^\circ, -7.7^\circ < b < 8.7^\circ$	Taylor et al. 2003
1.4 GHz	$\approx 1'$	$147.3^\circ > l > 74.2^\circ, -3.6^\circ < b < 5.6^\circ$	Taylor et al. 2003
1.4 GHz	$\approx 1'$	$67^\circ > l > 18^\circ, b \pm 1.5^\circ$	Stil et al. 2006
1.4 GHz (PI)	$\approx 1'$	$358^\circ > l > 253^\circ, b \pm 1.5^\circ$	Haverkorn et al. 2006
1.4 GHz (PI)	$\approx 1'$	$175^\circ > l > 66^\circ, -3^\circ < b < 5^\circ$	Landecker et al. 2010
1.4 GHz	$9'$	$162^\circ > l > 93^\circ, b \pm 4^\circ$	Kallas and Reich 1980
1.4 GHz	$9.4'$	$240^\circ > l > 95.5^\circ, -4^\circ < b < 5^\circ$	Reich et al. 1990a, 1997
1.4 GHz (PI)	$9'$	selected areas	Uyaniker et al. 1999
1.4 GHz	$\approx 1'$	$332.5^\circ > l > 325.5^\circ, -0.5^\circ < b < 3.5^\circ$	Gaensler et al. 2001
2.4 GHz	$10.4'$	$238^\circ > l > 365^\circ, b \pm 5^\circ$	Duncan et al. 1995
2.4 GHz (PI)	$10.4'$	$238^\circ > l > 5^\circ, b \pm 5^\circ$	Duncan et al. 1997
2.7 GHz	$4.3'$	$357.4^\circ < l < 76^\circ, b \pm 1.5^\circ$	Reich et al. 1984
2.7 GHz (PI)	$6'$	$74^\circ > l > 4^\circ.9, b \pm 1.5^\circ$	Junkes et al. 1987
2.7 GHz	$4.4'$	$76^\circ > l > 358^\circ, b \pm 5^\circ$	Reich et al. 1990b
2.7 GHz	$4.4'$	$240^\circ > l > 76^\circ, b \pm 5^\circ$	Fürst et al. 1990
2.7 GHz (PI)	$4.3'$	$74^\circ > l > 4^\circ.9, b \pm 5^\circ$	Duncan et al. 1999
5 GHz (PI)	$9'$	$129^\circ > l > 122^\circ, b \pm 5^\circ$	Sun et al. 2007
5 GHz (PI)	$9'$	$230^\circ > l > 129^\circ, b \pm 5^\circ$	Gao et al. 2010
5 GHz (PI)	$9'$	$122^\circ > l > 60^\circ, b \pm 5^\circ$	Xiao et al. 2011
5 GHz (PI)	$9'$	$60^\circ > l > 10^\circ, b \pm 5^\circ$	Sun et al. 2011

PI with polarization data



■ Fig. 13-11

A section of the Galactic plane at 1.4 GHz (Landecker et al. 2010) (reproduced with permission © ESO)

Uyanıker et al. 1999; Wolleben and Reich 2004). These are foreground clouds of diffuse thermal gas and magnetic fields which Faraday-rotate or Faraday-depolarize the extended polarized emission from the background. In addition to the well-known polarized SNRs and unpolarized HII regions, molecular clouds, pulsar-wind nebulae, and planetary nebulae were identified as Faraday Screens. Depending on the rotation angle and the polarization angle of the background emission, such screens may appear bright or dark. The strength and structure of regular fields can be estimated via the RM. Such observations can trace magnetic structures to sub-parsec scales.

The present data set on the intensity distribution and polarization of the Galactic plane (listed in [Table 13-4](#)) is impressive. Sensitive surveys at higher radio frequencies are needed to allow a systematic study of the Faraday Screen phenomenon. This is being achieved, e.g., by a Sino-German survey of the Galactic plane at 5 GHz with the Urumqi telescope (Gao et al. 2010; Sun et al. 2007, 2011; Xiao et al. 2011).

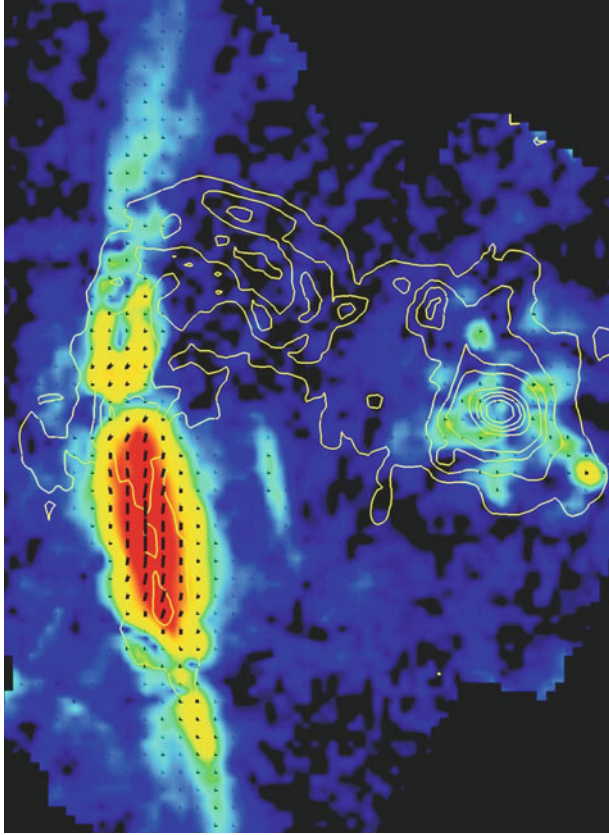
3.2.3 The Galactic Center

The Galactic Center is unique source with unusual radio continuum features. Mapping of the Galactic Center region by Yusef-Zadeh et al. (1984) showed several features vertical to the plane. The radio continuum emission is most intense and has a flat spectral index (Reich et al. 1988), for many years accounted to thermal emission. However, this intense emission is highly polarized (e.g., Seiradakis et al. 1985) and has been interpreted to be due to mono-energetic electrons (Lesch et al. 1988). Also, the polarization “strings” imply vertical magnetic structures, much different from the azimuthal directions of the magnetic fields seen along the Galactic plane. Mapping of the Galactic Center at 32 GHz (Reich 2003) showed that RMs in excess of $\pm 1,600 \text{ rad m}^{-2}$ are present in the vertical structures. The discussion about the intensity of the magnetic fields have yielded very high values (mG range) based on the high rotation measures. Zeeman splitting observations by Yusef-Zadeh et al. (1996) also suggest mG magnetic fields. Other authors (e.g., Crocker et al. 2010) suggest much lower values in the 50–100 μG range.

Detailed high-resolution studies also brought controversial results. High resolution maps of the Galactic Center (e.g., Nord et al. 2004) showed a spiral structure at the position of Sgr A* and thin vertical radio continuum “strings.” Polarimetric observations at sub-mm wavelengths suggest a stretched magnetic field (Novak et al. 2000), as expected in sheared clouds. A recent interpretation of the magnetic field phenomena in the Galactic Center was given by Ferrière (2009). If our Galaxy does not differ much from nearby galaxies, the vertical field detected close to the center is a local phenomenon ([Fig. 13-12](#)).

3.3 Faraday Rotation of Extragalactic Radio Sources and Pulsars

Faraday rotation (FR) is a powerful tool for studying magnetic fields. First, ionospheric rotation, later Faraday effects due to the Galactic ISM were detected soon after the discovery of linear polarization of the Galactic radio waves. At first, the FR of diffuse emission was studied. Later, with increasing samples of EGRS, modeling of the magnetic field was attempted. Finally, pulsars, most of which are concentrated to the Galactic plane, were used to model the Galactic magnetic fields.



■ Fig. 13-12

Galactic Center region. Total intensity (*contours*), polarized intensity (*colors*), and B-vectors at 9 mm, observed with the Effelsberg telescope. The map size is about $23' \times 31'$ along Galactic longitude and latitude. The Galactic Center is located at the peak of total emission (from Wolfgang Reich, MPIfR)

3.3.1 Extragalactic Radio Sources (EGRS)

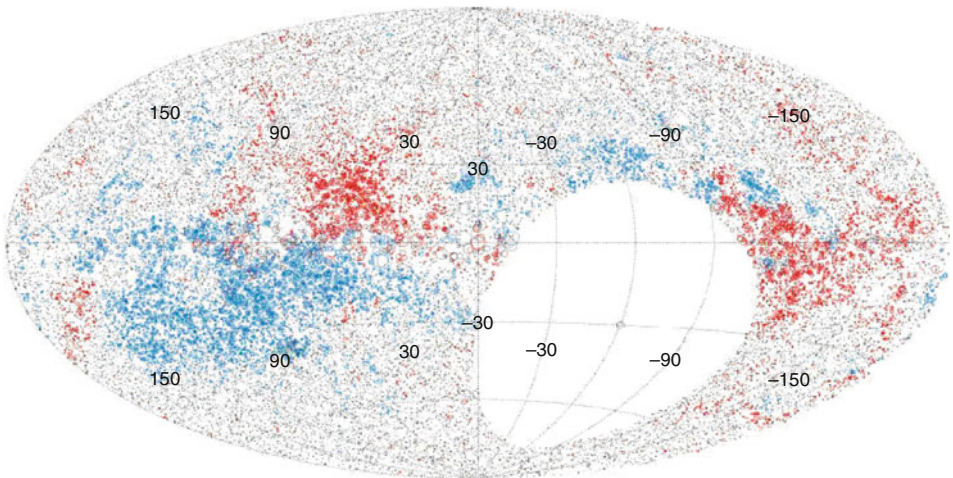
Faraday rotation measures (RMs) toward EGRS originate in the source itself and in the magneto-ionic media in the foreground (intergalactic space, intervening galaxies, Milky Way, interplanetary space and ionosphere of the Earth). The contributions from intergalactic space, intervening galaxies, and interplanetary space are generally small. The contribution from the ionosphere of the Earth is subtracted with help of calibration sources with known polarization angle, leaving RM from the Milky Way and intrinsic RM.

Here, a word of caution must be given. The intrinsic polarization and RM of any EGRS may originate in the nucleus of a radio galaxy or in the extended lobes. Hence, when making observations at various frequencies to obtain the correct RM, care must be taken that the same source

structure is measured. In particular, there are problems in combining data from single-dish observations with those of an interferometer at other frequencies. More recent observations use many adjacent frequency channels to accurately determine the RM if a sufficiently wide band is used (*RM Synthesis*, ● Sect. 2.4). This also helps to separate the intrinsic RM from that in the foreground. If RM Synthesis is not available, averaging over a large number of RMs is used to reduce the intrinsic contributions.

The earliest catalogues of RM toward EGRS were collected by Simard-Normandin and Kronberg (1980), showing the all-sky distribution of RMs. In this compendium of sources, there were only a few sources with measured RM along the Galactic plane, where the Galactic magnetic fields are concentrated, as was seen in the all-sky continuum surveys. Hence, the interpretation of this data gave us an indication of a local magnetic field only. In recent years, additional data on the RM of sources in the Galactic plane were obtained (Brown et al. 2007; Van Eck et al. 2011). However, all these surveys cover only partially the Galactic plane so that interpretation is difficult. The highest observed values were $|RM| \approx 1,000 \text{ rad m}^{-2}$ toward the Galactic Center. Similarly high $|RM|$ values were determined by Roy et al. (2008), who surveyed an area directly at the center of our Galaxy. There is neither a uniform coverage of the Galactic plane nor of the whole sky as yet. A statistical method to visualize the RM distribution over the sky was developed by Johnston-Hollitt et al. (2003) who used 800 sources. This work showed several areas of consistent RM values (of the same sign) as well as structures above and below the plane.

The data for the southern sky is still very sparse and needs to be extended. The most recent addition to the data set was undertaken by Taylor et al. (2009), who reanalyzed the NRAO VLA Sky Survey (NVSS) (● Fig. 13-13). This study involved 37,543 sources and added a huge number of new RMs, but is limited by the rather close frequency separation of the two frequency bands



■ Fig. 13-13

Rotation measures determined from the NVSS catalog (Taylor et al. 2009) (reproduced by permission of the AAS)

which leads to large RM errors. The averaged RM toward extragalactic sources reveal no large-scale reversal across the plane around Galactic longitudes 120° and also -120° (according other observations showing a reversed sign in the southern region, not shown in [Fig. 13-13](#)): The local disk field is part of a large-scale symmetric field structure. However, toward the inner Galaxy the RM signs are opposite above and below the plane. This reversal may be due to local features (Wolleben et al. 2010b) or to an antisymmetric toroidal field in the Milky Way's halo ([Sect. 3.5](#)).

Another project to increase the number of RMs over the whole sky was undertaken at the Effelsberg radio telescope, combining polarization data in eight channels in two bands around 1.4 and 1.6 GHz. This instrumental combination allows for accurate determination of the RM of sources. Some 1,600 new RMs were added, and a preliminary result is given in Wielebinski et al. (2008). The preliminary data for 2,469 sources were used to model the Galactic magnetic field (Sun et al. 2008; [Sect. 3.5](#)).

3.3.2 Pulsars

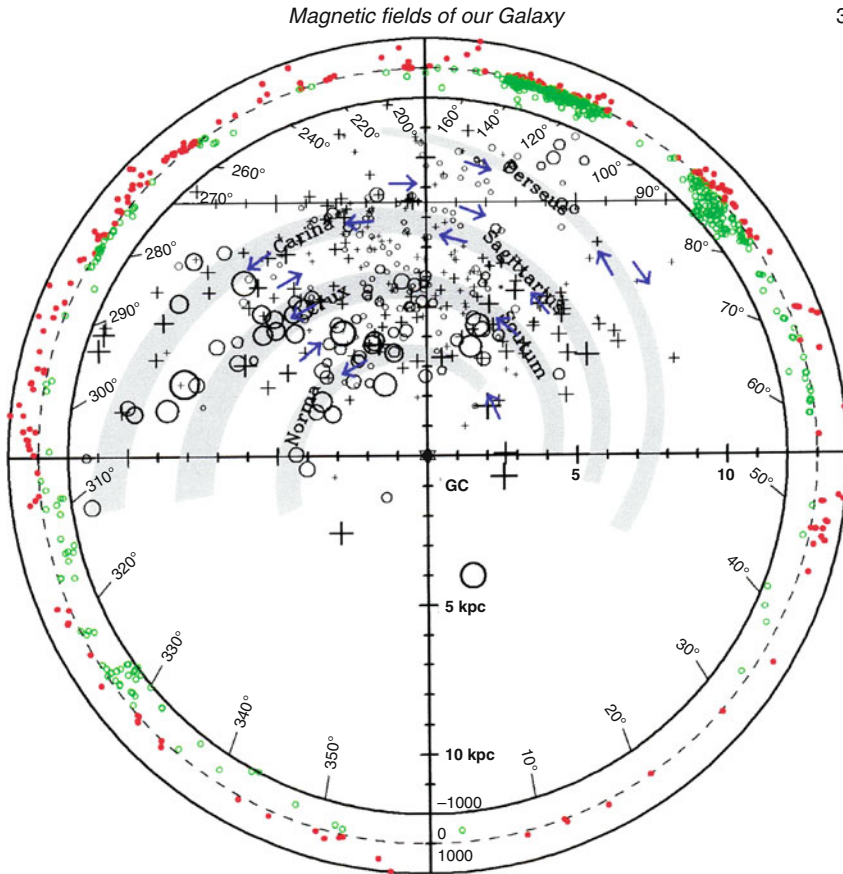
Pulsars are the ideal sources to probe the magnetic fields through the Faraday effect. Since pulsars have no measurable angular structure and they are highly polarized, they are the ideal probes. Pulsars are Galactic objects and hence, their distribution is close to the Galactic plane toward the inner Galaxy. In fact, very few pulsars are known toward the anticenter of the Galaxy. Hence, a combination of pulsars and EGRS is optimal for studies of the Galactic magnetic field. Pulsars also allow measurement of the Dispersion Measure (DM) which follows from the signal delay occurring in the foreground medium. Together with the RM, the value of the average regular magnetic field in the line of sight can be deduced:

$$\langle B_{\parallel} \rangle = 1.232 \frac{\overline{\text{RM}}}{\text{DM}} \mu\text{G}$$

Application gives an average strength of the regular field in the local spiral arm of $1.4 \pm 0.2 \mu\text{G}$. In the inner Norma arm, the average strength of the regular field is $4.4 \pm 0.9 \mu\text{G}$. However, this estimate is only valid if variations in the regular strength and in electron density are *not correlated*. If they are correlated, the above formula gives an overestimate of $\langle B_{\parallel} \rangle$ and an underestimate for anti-correlated variations (Beck et al. 2003). The dispersion of pulsar RMs yields an estimate for the turbulent field strength of about $5 \mu\text{G}$ and for the turbulence length of about 50 pc (Rand and Kulkarni 1989).

The major compilation of pulsar rotation measures, also using already published data, are given in Han et al. (2006, 2009). Additional results are found in Mitra et al. (2003), Noutsos et al. (2008) and Van Eck et al. (2011). The distribution of rotation measures, as given by Han (2008), shows a huge variation of signs and magnitudes. This may indicate a large-scale regular magnetic field with multiple reversals ([Sect. 3.5](#)) or the effect of localized regions, e.g., HII regions (Mitra et al. 2003).

The RM values increase for distant objects, but very few pulsars were found beyond the Galactic Center ([Fig. 13-14](#)). The limit of $|\text{RM}| \approx 1,000 \text{ rad m}^{-2}$ for EGRS holds also for pulsars. This seems to indicate that the RM toward EGRS is partly averaged out in passage through the Galaxy. The large-scale regular field of the outer Milky Way is either weak or frequently reversing its direction.

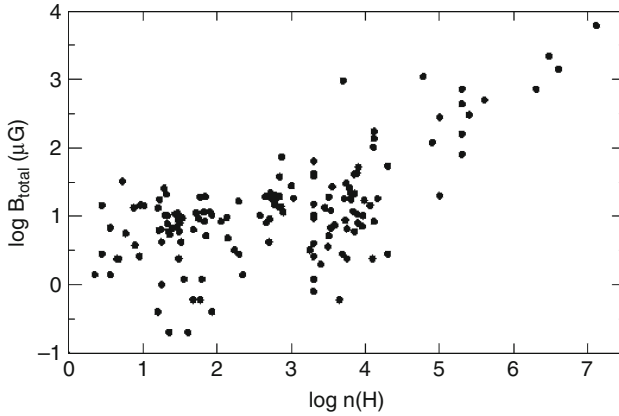


■ Fig. 13-14

Faraday rotation measures (RM) of pulsars in the Milky Way (within *inside circle*) and of extragalactic radio sources (between *inside and outside circles*). *Plus signs* indicate positive RM toward pulsars, *small circles* negative RM. *Red symbols* indicate positive RM toward extragalactic sources, *green symbols* negative RM. The *blue arrows* suggest large-scale magnetic fields along a model of spiral arms in the Milky Way. Our sun is located at the upper crossing of coordinate lines (Han 2008)

3.4 Zeeman Effect

The Zeeman effect is the most direct method of measuring magnetic fields. It has been used in the optical range for detecting magnetic fields in the Sun and in stars. At radio wavelength the use of the Zeeman effect proved to be more difficult. For one, the frequency shifts caused by the weak magnetic fields are minute and require sophisticated instrumentation. The HI line gave the first definitive detections, usually in absorption toward strong Galactic sources (Verschuur 1968). The technique was refined so that at present magnetic fields as weak as $\approx 5 \mu\text{G}$ can be detected with the Arecibo telescopes. The observation of the Zeeman effect in the OH molecule (e.g., Crutcher et al. 1987) advanced the field further.



■ Fig. 13-15

Zeeman measurements of the total magnetic field in gas clouds plotted against the hydrogen volume density n_{H} (in cm^{-3}). To derive the total field B_{total} , each measured line-of-sight component was multiplied by a factor of 2 which is the average correction factor for a large sample (Crutcher et al. 2010)

It became clear that many of the positive detections were in molecular clouds with maser sources. Strong magnetic fields (≈ 80 mG) were detected in interstellar H_2O maser clouds (Fiebig and Güsten 1989). Millimeter-wavelength astronomy gave us additional results for high recombination lines (Thum and Morris 1999) or in such molecules as CN (Crutcher et al. 1999) or CCS (Levin et al. 2000). A compilation of present-day Zeeman measurements of the magnetic field in gas clouds (● Fig. 13-15) gives a mean total field in the cold neutral interstellar gas of $6 \pm 2 \mu\text{G}$ so that the magnetic field dominates thermal motion but is in equipartition with turbulence, as also found on much larger scales in external galaxies (▶ Sect. 4.2). Beyond cloud densities of $\approx 1,000 \text{ cm}^{-3}$, the field strength scales with $n^{0.65 \pm 0.05}$.

The importance of magnetic fields in the star-formation process is obvious. Diffuse clouds are subcritical with respect to collapse and probably balanced by magnetic fields, while dense molecular are supercritical and collapse. The transition from subcritical to supercritical state may be the result of ambipolar diffusion or turbulence. Zeeman observations in the HI and OH lines can measure the ratio of mass to magnetic flux in the cloud envelope and the core. A smaller ratio in the core may indicate that supersonic turbulence plays a similarly important role as ambipolar diffusion (Crutcher et al. 2009), but effects of the field geometry also have to be taken into account (Mouschovias and Tassis 2009). More and higher-quality data are needed.

The use of the Zeeman data for the investigation of a large-scale regular magnetic field of the Galaxy was attempted by several authors (e.g., Fish et al. 2003). The number of detected sources was rather small and the interpretation in terms of Galactic magnetic fields rather inconclusive. Han and Zhang (2007) collected a large data set of Zeeman results and studied the question if the magnetic fields in molecular clouds preserve information of the direction of the large-scale magnetic fields in the spiral arms. In spite of a larger data set, all that the conclusion offered was that clouds “may still remember the directions of regular magnetic fields in the Galactic ISM to some extent.”

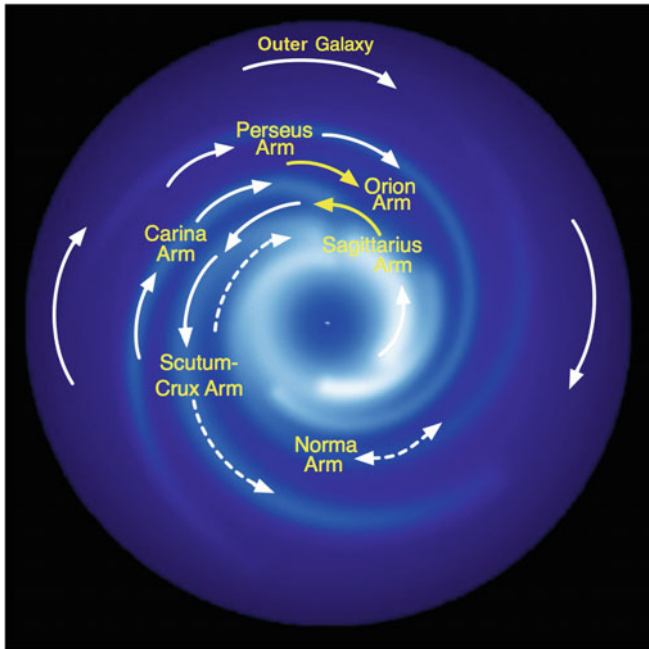
3.5 Modeling the Magnetic Field of the Milky Way

Based on all the data described in previous sections models of the magnetic fields of the Milky Way have been repeatedly made. At first, the low-frequency all-sky data was used to describe the Galactic nonthermal emission (e.g., Yates 1968) produced in magnetic fields. The all-sky survey of Haslam et al. (1982) has been interpreted by Phillips et al. (1981) and Beuermann et al. (1985). Using the data on the HII regions (e.g., Georgelin and Georgelin 1976) of the Galaxy, it could be shown that the spiral structure is also seen in the diffuse radio continuum emission.

The RM data first for pulsars and later for extragalactic radio sources (EGRS) also led to modeling of the magnetic fields of the Milky Way. The RMs toward pulsars are due to the ISM in the direction of the inner Galaxy. There are few pulsars known outside the inner quadrants. The data on EGRS gave information about the Faraday effects over much of the sky but not for the Galactic plane. Until recently, there were very few EGRS observed through the inner Galactic plane, but much more data are now available.

The collection of 543 rotation measures of EGRS distributed across the sky by Simard-Normandin and Kronberg (1980) showed that there were areas with similar RM directions, suggesting organized magnetic fields over larger Galactic scales. Pulsar observers were the first to point out that in addition to large areas of similar magnetic field orientations, there were some regions where the field *reverses* along Galactic radius. These results were analyzed with wavelets (e.g., Stepanov et al. 2002) and confirmed the existence of at least one large-scale reversal. Since most of the EGRS investigated were away from the Galactic plane, they did not trace the Galactic magnetic field in the disk, more likely were indications of some local magnetic features. The number of RMs has been steadily increasing (e.g., Taylor et al. 2009) which increased the sampling of the Galaxy considerably. The same general conclusions were reached as in the earlier work – organized magnetic structures in sections of the Galaxy and highly disorganized magnetic fields toward the central region. One of the disadvantages of the available data is the fact that the southern sky data is very spare and needs additional observations. Progress has been made in observations of EGRS in the Galactic plane (Brown et al. 2007; Van Eck et al. 2011), but the sampling is still not uniform and not dense enough along the plane. The RM of EGRS in the very center of the Galaxy was studied by Roy et al. (2008) who detected mainly positive RM values, suggesting a magnetic field aligned with a central bar.

The analysis of data from radio continuum all-sky surveys at 1.4 and 23 GHz, from RMs toward EGRS, the best available thermal electron model and an assumed cosmic ray distribution (Sun and Reich 2010; Sun et al. 2008) constrained the average field strength of the Galaxy to $\approx 2 \mu\text{G}$ for the regular field and $\approx 3 \mu\text{G}$ for the random field in the solar neighborhood, similar to the results from pulsar RMs (☛ Sect. 3.2.2). An axisymmetric (ASS) magnetic field configuration (☛ Sect. 2.6) fits the observed data best, but one large-scale reversal is required in the Sagittarius Arm about 1–2 kpc inside the solar radius (☛ Fig. 13-16). The local field is oriented parallel to the plane and its direction is symmetric (even parity) with respect to the Galactic plane, while the toroidal component of the halo field has different directions above and below the Galactic plane (odd parity, see ☛ Sect. 2.6), to account for the different signs of the observed RM data. If this antisymmetry is globally valid for the Milky Way, its halo field has a dipolar pattern, in contrast to that found in external galaxies (☛ Sect. 4.7). However, some of the asymmetry can be explained by distorted field lines around a local HI bubble (Wolleben et al. 2010b). Future modeling should take into account all the existing data, and future observations with better sampling of the sky are needed (☛ Sect. 5).



■ Fig. 13-16

Model of the magnetic field in the Milky Way, derived from Faraday rotation measures of pulsars and extragalactic sources. Yellow arrows indicate confirmed results, while white and dashed arrows still need confirmation (Van Eck et al. 2011)

Pulsars are ideal objects to deduce the Galactic magnetic field. Since most pulsars are concentrated along the Galactic plane, they sample the field in the disk. The pulsar and EGRS data led to several attempts to model the Galactic magnetic field (e.g., Han et al. 2006; Manchester 1974; Vallée 2005; Van Eck et al. 2011).

Following a suggestion of Rand and Kulkarni (1989), the analysis has concentrated on attempts to fit either a bisymmetric spiral (BSS) or an axisymmetric spiral (ASS) field structure to the existing data. The local magnetic field in the Perseus arm is clockwise. A magnetic field reversal seems to be present toward the Sagittarius arm ($l \approx 50^\circ$) that was often used as an argument for a BSS field structure, although such a reversal can be local or be part of a more complicated field structure. Detailed analysis (e.g., Noutsos et al. 2008; Vallée 1996) has shown that this concept of a single large-scale field mode is not compatible with the data. The analysis of the previous interpretations by Men et al. (2008) also showed that, presently, there is no definite proof of either the BSS or the ASS configuration.

Studies of the effects of large HII regions on the RM changes of pulsars beyond these Faraday screens showed that some earlier interpretations that some of the claimed field reversals are only local (Mitra et al. 2003). Furthermore, the comparison of the RM of pulsar and EGRS toward the Galactic Center (Brown et al. 2007) revealed similar values of RM, as if there were no other half of the Galaxy. This result suggests that the RMs are dominated by local ISM features and that the large-scale field is weak and cannot be delineated from the available data. Only RM data free from the effects of HII regions should be used, as demonstrated by Nota and Katgert (2010).

Statistically safe is the existence of one large-scale field reversal in the Milky Way, which is puzzling. Very few large-scale reversals have been detected so far in external spiral galaxies, and none along the radial direction (▶ Sect. 4.4). The different observational methods may be responsible for this discrepancy between Galactic and extragalactic results. RMs in external galaxies are averages over the line of sight through the whole disk and halo and over a large volume traced by the telescope beam, and they may miss field reversals, e.g., if these are restricted to a thin region near to the galaxy plane. On the other hand, the results in the Milky Way are based on RMs of pulsars, which trace the magneto-ionic medium near the plane. Alternatively, the Milky Way may be “magnetically young” and may still not have generated a coherent large-scale field over the whole disk. The timescale for fully coherent fields can be longer than the galaxy age, e.g., if frequent interactions with other galaxies occur (▶ Sect. 2.6). The observed field reversal may be a temporary phenomenon with a limited lifetime. The magnetic field structure of the Milky Way is probably quite complex and shows details which cannot be resolved yet in external spiral galaxies.

Little is known about the large-scale field in the Milky Way’s halo. From a survey of RMs of EGRS toward the Galactic poles, Mao et al. (2010) derived a local field perpendicular to the plane of $+0.31 \pm 0.03 \mu\text{G}$ toward the south Galactic pole, but no significant field toward the north Galactic pole. This is neither consistent with the dipolar halo field as suggested from the antisymmetry of the toroidal field nor with a quadrupole halo field as found in several external galaxies (▶ Sect. 4.7) – adding another puzzle. In the Galactic Center, vertical magnetic fields exist which apparently extend into the halo. Again, the halo field may be more complicated than predicted by mean-field dynamo models, and regions with different field directions may exist.

While observations in the Milky Way can trace magnetic structures to much smaller scales than in external galaxies, the large-scale field is much more difficult to measure in the Milky Way. This information gap will be closed with future radio telescopes which will find many new pulsars in the Milky Way (and in nearby galaxies) and which allow us to observe the detailed magnetic field structure also in external galaxies (▶ Sect. 5).

4 Galaxies

Magnetic fields in external galaxies can be observed with the same methods as in the Milky Way, except for extragalactic pulsars which have been found so far only in the Magellanic clouds. Naturally, the spatial resolution of the telescopes is much worse in galaxies, and the detailed structure of extragalactic fields on scales below about 100 pc is still invisible. On the other hand, the large-scale field properties, like the overall pattern and the total extent, can be best measured in external galaxies. Observations in the Milky Way and in external galaxies are complementary.

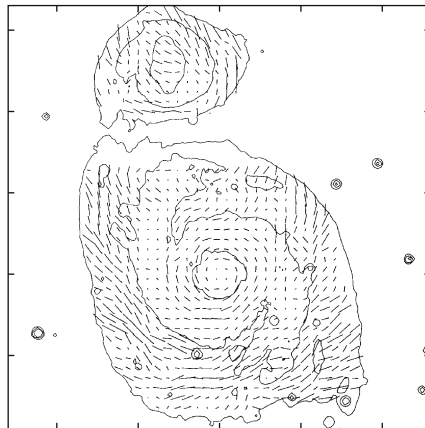
4.1 Optical Polarization, Infrared Polarization, and Zeeman Effect

Weak optical polarization (generally below 1%) is the result of extinction by elongated dust grains in the line of sight which are aligned in the interstellar magnetic field (the Davis–Greenstein effect, see ▶ Sect. 2.1). Optical polarization surveys yielded the large-scale structure of the field in the local spiral arm of our Milky Way (▶ Sect. 3). The first extragalactic results by Hiltner (1958) were based on starlight polarization of globular clusters in M31 and showed that

the magnetic field is aligned along the galaxy's major axis. Polarization of starlight in the LMC also gave evidence for ordered fields near 30 Dor (Mathewson and Ford 1970b) and possibly along the Magellanic stream (Schmidt 1976).

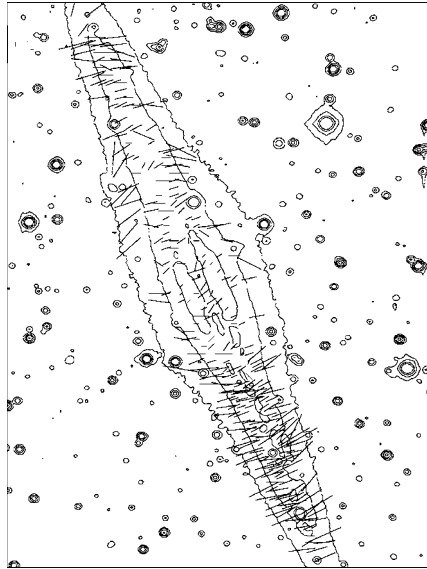
Polarization from diffuse optical light was used to search for large-scale magnetic fields, though some unknown fraction of the polarized light is due to scattering on dust particles. Indications for ordered fields along the spiral arms were found in M82 (Elvius 1962), M51 and M81 (Appenzeller 1967; Scarrott et al. 1987), in NGC1068 (Scarrott et al. 1991), and in NGC6946 (Fendt et al. 1998). The pattern in M51 (◀ Fig. 13-17) agrees well with the radio polarization results (see ▶ Fig. 13-23) in the inner spiral arms, but large differences are seen in the outer arms and in the companion galaxy, which is unpolarized in the radio image. In the Sa-type edge-on Sombrero galaxy M104 and the Sb-type edge-on NGC4545, optical polarization indicates a field along the prominent dust lane and vertical fields above the plane (Scarrott et al. 1990), in agreement with the results from radio polarization (▶ Sect. 4.7). The polarization of the Sc-type edge-on galaxies NGC891 (▶ Fig. 13-18), NGC5907, and NGC7331 shows fields near the galaxy plane which are predominantly oriented perpendicular to the plane (Fendt et al. 1996), possibly aligned along the vertical dust filaments observed in these galaxies. Radio continuum polarization, on the other hand, traces the magnetic fields in the diffuse medium which are mostly oriented parallel to the plane and have significant vertical components only beyond some height above the plane (▶ Sect. 4.7).

Correction of the diffuse optical polarization for scattering effects is difficult and has never been attempted so far. Instead, polarization techniques were developed in the infrared where scattering is negligible. Near-IR polarization in a dust lane of the edge-on galaxy NGC4565 indicates a plane-parallel field (Jones 1989), similar that seen at radio wavelengths. In the far-IR and sub-mm ranges, the emission of aligned dust grains is intrinsically polarized and the degrees of polarization can reach several percent. The galaxy M82 was observed with the JCMT polarimeter at 850 μm (Greaves et al. 2000), but the derived bubble-type field pattern is in contrast to the radio data indicating a field that is oriented radially outward (Reuter et al. 1992).



■ Fig. 13-17

Spiral galaxy M51. E-vectors of the optical polarization of diffuse light which trace the spiral magnetic field orientation (Scarrott et al. 1987). Compare with the radio polarization map in ▶ Fig. 13-23



■ Fig. 13-18

Edge-on spiral galaxy NGC891. E-vectors of optical polarization of the diffuse light, indicating vertical magnetic fields (Fendt et al. 1996). Compare with the radio polarization map in [▶ Fig. 13-41](#) (reproduced with permission © ESO)

Potential differences between IR, sub-mm, and radio polarization data should be investigated with the forthcoming polarimeters at the JCMT, APEX, ALMA, and SOFIA telescopes.

Zeeman measurements in external galaxies are still very rare. Robishaw et al. (2008) detected the effect in the OH megamaser line at 18 cm in 5 distant starburst galaxies and derived field strengths in these dense gas clouds between 0.5 and 18 mG. Measurements in nearby galaxies will become possible with the Square Kilometre Array ([▶ Sect. 5](#)).

4.2 Magnetic Field Strengths

The dynamical importance of the total magnetic field B may be estimated by its energy density which is proportional to B^2 . Due to its vector nature, the dynamical effect of the magnetic field also depends on its structure and degree of ordering ([▶ Sect. 4.4](#)). The average strength of the component B_{\perp} of the total field and $B_{\text{ord},\perp}$ of the resolved ordered field in the plane of the sky can be derived from the total and polarized radio synchrotron intensity, respectively, if energy-density equipartition between total cosmic rays and total magnetic field B is valid ([▶ Sect. 2.2](#)). The field strengths B_{\perp} are given by the mean surface brightnesses (intensities) of the synchrotron emission, hence depend neither on the size nor on the distance of the galaxy.

The observed radio emission from galaxies has a contribution of thermal emission from ionized gas (and at frequencies beyond about 50 GHz also from dust) which needs to be subtracted to obtain the pure synchrotron part. The mean thermal fraction is about 10% at 21 cm

and about 30% at 3 cm, but may increase to $\geq 50\%$ in star-forming regions. A proper subtraction of the radio thermal intensity needs an independent thermal template, e.g., the $H\alpha$ intensity corrected for extinction with help of a dust model based on far-infrared data (Tabatabaei et al. 2007). For a crude separation of thermal and synchrotron intensity components, comparison of the observed radio spectral index with an assumed synchrotron spectral index is sufficient.

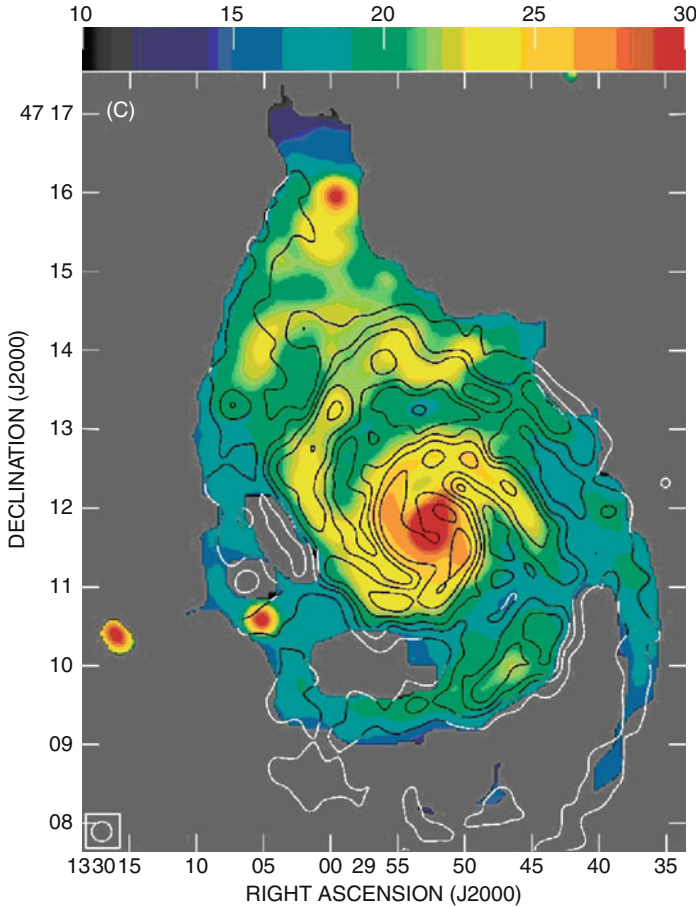
The average equipartition strength of the total field (corrected for inclination) for a sample of 74 spiral galaxies is $B = 9 \pm 2 \mu\text{G}$ (Niklas 1995). Dwarf galaxies host fields of similar strength as spirals if their star-formation rate per volume is similarly high. Blue compact dwarf galaxies are radio bright with equipartition field strengths of 10–20 μG (Klein et al. 1991). Spirals with moderate star-forming activity and moderate radio surface brightness like M31 (Fig. 13-26) and M33 (Fig. 13-36), our Milky Way's neighbors, have $B \approx 6 \mu\text{G}$. In “grand-design” galaxies with massive star formation like M51 (Fig. 13-23), M83 (Fig. 13-24), and NGC6946 (Fig. 13-25), $B \approx 15 \mu\text{G}$ is a typical average strength of the total field.

In the density-wave spiral arms of M51, the total field strength B is 25–30 μG (Fig. 13-19). Field compression by external forces like interaction with other galaxies may amplify the fields (Sect. 4.8). The strongest fields in spiral galaxies (50–300 μG) are found in starburst galaxies like M82 (Klein et al. 1988), the “Antennae” NGC4038/9 (Fig. 13-44), in nuclear starburst regions, like in the nuclear rings of NGC1097 (Fig. 13-35) and other barred galaxies, and in nuclear jets (Fig. 13-49).

If energy losses of cosmic-ray electrons are significant in starburst regions or massive spiral arms, the equipartition values are lower limits (Sect. 2.2). The average equipartition field strength in normal spirals is proportional to the average gas surface density, but this relation is no longer valid for starburst galaxies (Thompson et al. 2006). Due to strong energy losses of the cosmic-ray electrons and even protons, the equipartition field strength is probably underestimated by a factor of a few. Field strengths of 0.5–18 mG were detected in starburst galaxies by the Zeeman effect in the OH megamaser emission line at 18 cm (Robishaw et al. 2008). However, these values refer to highly compressed gas clouds and are not typical for the interstellar medium.

The relative importance of various competing forces in the interstellar medium can be estimated by comparing the corresponding energy densities. In the local Milky Way, the energy densities of the stellar radiation field, turbulent gas motions, cosmic rays, and total magnetic fields are similar (Boulares and Cox 1990). The mean energy densities of the total magnetic field and of the cosmic rays in NGC6946 and M33 are $\approx 10^{-11}$ and $\approx 10^{-12}$ erg cm^{-3} , respectively (Beck 2007; Tabatabaei et al. 2008), similar to that of the turbulent motions of the cold, neutral gas with density ρ across the star-forming disk (Fig. 13-20). The turbulent energy may be underestimated if V_{turb} is larger than 7 km s^{-1} or if the warm gas also contributes. The energy density of the warm ionized gas E_{th} with electron density n_e is one order of magnitude smaller than that of the total magnetic field E_B , which means that the ISM in spiral galaxies is a *low- β plasma* ($\beta = E_{\text{th}}/E_B$), similar to that of the Milky Way (Boulares and Cox 1990). Hot gas also contributes to E_{th} , but its contribution is small. The overall dominance of turbulent energy is surprising because the supersonic turbulence should heat the gas but is also derived from numerical ISM simulations (de Avillez and Breitschwerdt 2005). Further investigations with higher resolution are needed.

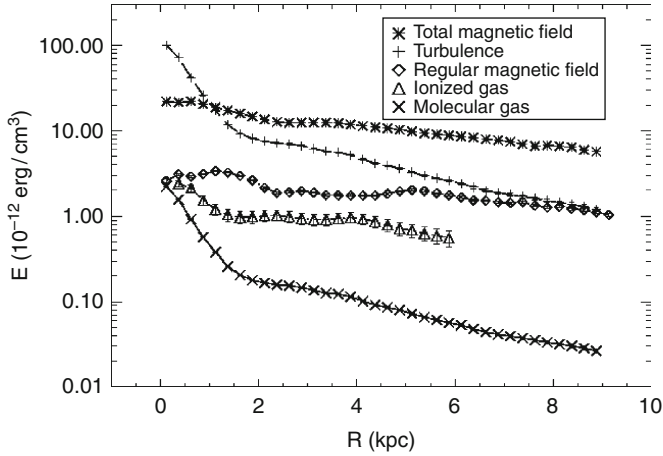
The radial distribution of synchrotron intensity in many spiral galaxies is well described by an exponential decrease with a scalelength l_{syn} of about 4 kpc. In case of equipartition between the energy densities of magnetic fields and cosmic rays, the scalelength of the field is $(3 - \alpha) l_{\text{syn}} \approx 16$ kpc (where $\alpha \approx -1$ is the synchrotron spectral index). The scalelength



■ Fig. 13-19

Spiral galaxy M51. Total equipartition magnetic field strengths (in μG), corrected for the inclination of the galaxy (Fletcher et al. 2011)

of the regular field is even larger (► Fig. 13-20). These are still lower limits because energy losses of cosmic-ray electrons increase with increasing distance from their origin in the galaxy's star-forming regions, and a lower density of cosmic-ray electrons needs a stronger field to explain the observed synchrotron intensity. Fields in the outer disk of galaxies can be amplified by the mean-field dynamo even without star-formation activity because turbulence can be generated by the magneto-rotational instability (MRI, ► Sect. 2.6). The typical scalelengths of the density of neutral and ionized gas are only about 3 kpc so that the magnetic field energy dominates over the turbulent energy in the outer region of galaxies if a constant turbulent velocity is assumed (► Fig. 13-20). The speculation that magnetic fields may affect the global gas rotation (Battaner and Florido 2000) needs testing by future radio observations with higher sensitivity and at low frequencies where energy loss of electrons is smaller.



■ Fig. 13-20

Spiral galaxy NGC6946. Radial variation of the energy densities of the total magnetic field E_B ($B^2/8\pi$), ordered (mostly regular) magnetic field ($B_{\text{reg}}^2/8\pi$), turbulent motion of the neutral gas E_{turb} ($0.5 \rho_n V_{\text{turb}}^2$, where $V_{\text{turb}} \approx 7 \text{ km s}^{-1}$), thermal energy of the ionized gas E_{th} ($0.5 n_e k T_e$), and thermal energy of the molecular gas E_n ($0.5 \rho_n k T_n$), determined from observations of synchrotron and thermal radio continuum, and the CO and HI line emissions (Beck 2007) (reproduced with permission © ESO)

In spiral arms of galaxies, the typical degree of radio polarization is only a few percent. The total field B_{\perp} in the spiral arms must be mostly *turbulent* with random orientations within the telescope beam, which typically corresponds to a few 100 pc at the distance of nearby galaxies. The typical ratios of turbulent fields to resolved ordered fields are ≥ 5 in spiral arms and circumnuclear starburst regions, 0.5–2 in interarm regions, and 1–3 in radio halos. Turbulent fields in spiral arms are probably generated by turbulent gas motions due to supernovae (de Avillez and Breitschwerdt 2005), turbulence induced by spiral shocks (Dobbs and Price 2008), or the small-scale dynamo (◆ Sect. 2.6).

Magnetic turbulence occurs over a large spectrum of scales. The maximum scale of the turbulence spectrum in the Milky Way derived from the dispersion of rotation measures of pulsars is $d \approx 50 \text{ pc}$ (Rand and Kulkarni 1989). This scale can also be derived from the depolarization by the superposition of emission from turbulent fields at centimeter wavelengths (◆ Sect. 2.3). For a typical degree of polarization of 1% in spiral arms, 500 pc resolution in nearby galaxies and 1 kpc pathlength through the turbulent medium, $d \approx 40 \text{ pc } f^{1/3}$ where f is the filling factor of the ionized medium. At decimeter radio wavelengths, the same turbulent field causes Faraday dispersion (◆ Sect. 2.4). Typical depolarization of 50% at 20 cm, an average electron density of the thermal gas of 0.03 cm^{-3} and an average strength of the turbulent field of $10 \mu\text{G}$ yields $d \approx 10 \text{ pc } f^{-1}$. The two estimates agree for $d \approx 30 \text{ pc}$ and $f \approx 0.3$, consistent with the results derived with other methods.

Faraday dispersion can also be used to measure the strength of turbulent magnetic fields. However, the achievable accuracy is limited because the ionized gas density has to be determined from independent measurements. The increase of the mean degree of polarization at

20 cm with increasing distance from the plane of edge-on galaxies can constrain the parameters and, for NGC891 and NGC4631, yields strengths of turbulent magnetic fields in the plane of 11 μG and 7 μG and scale heights of 0.9 and 1.3 kpc, respectively (Hummel et al. 1991).

The strength of the *resolved ordered* (regular and/or anisotropic) fields B_{ord} in spiral galaxies is determined from the total equipartition field strength and the degree of polarization of the synchrotron emission. Present-day observations with typical spatial resolutions of a few 100 pc give average values of 1–5 μG . The ordered field is generally strongest in the regions between the optical spiral arms with peaks of about 12 μG , e.g., in NGC6946, is oriented parallel to the adjacent optical spiral arms, and is stronger than the tangled field. In several galaxies, like in NGC6946, the field forms coherent *magnetic arms* between the optical arms (► Fig. 13-25). These are seen at all wavelengths and hence cannot be the effect of weak Faraday depolarization in the interarm regions. Magnetic arms are probably signatures of higher modes generated by the mean-field dynamo (► Sect. 4.4). In galaxies with strong density waves, some of the ordered field is concentrated at the inner edge of the spiral arms, e.g., in M51 (► Fig. 13-23), but the arm–interarm contrast of the ordered field is small, much less than that of the turbulent field. The ordered field is more smoothly distributed in galaxies.

The *regular* (coherent) component of the ordered field can in principle be determined from Faraday rotation measures (► Sect. 2.4) if the mean electron density is known. In the Milky Way, the pulsar dispersion measure is a good measure of the total electron content along the pathlength to the pulsar. Only 19 extragalactic radio pulsars have been found so far, all in the LMC and SMC. In all other galaxies, the only source of information on electron densities of the warm ionized medium comes from thermal emission, e.g., in the $\text{H}\alpha$ line. However, thermal emission is dominated by the HII regions which have a small volume filling factor, while Faraday rotation is dominated by the diffuse ionized emission with a much larger filling factor. If the average electron density of the diffuse ionized medium in the Milky Way of 0.03–0.05 cm^{-3} is assumed also for other galaxies, Faraday rotation measures yield regular field strengths of a few μG . The strongest regular field of 8 μG was found in NGC6946 (Beck 2007), similar to the strength of the ordered field, hence most of the ordered field is regular in this galaxy. The similarity between the average regular (RM-based) and the ordered (equipartition-based) field strengths in NGC6946 and several other galaxies demonstrates that *both methods are reliable*, and hence no major deviations from equipartition occur in this galaxy on scales of a few kpc (but deviations may occur locally).

The situation is different in radio-bright galaxies like M51, where the average regular field strength is several times smaller than the ordered field (► Sect. 4.4). The total field is strong so that the energy loss of cosmic-ray electrons is high and the equipartition field is probably underestimated (► Sect. 2.2). This even increases the discrepancy between the two methods because the RM is not affected. The high-resolution observations of M51 indicate that anisotropic fields related to the strong density waves contribute mostly to the ordered field.

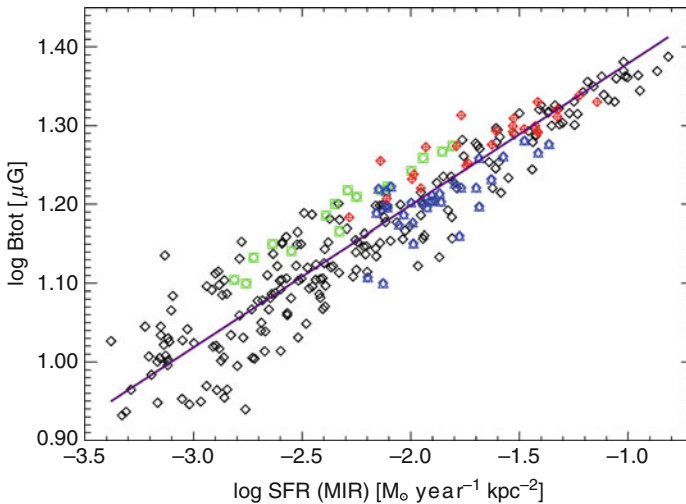
4.3 The Radio–Infrared Correlation

The highest total radio intensity (tracing the total, mostly turbulent field) generally coincides with the strongest emission from dust and gas in the spiral arms. The total radio and far-infrared (FIR) or mid-IR (MIR) intensities are highly correlated within galaxies. The exponent of the correlation in M51 is different in the central region, the spiral arms, and the interarm regions

(Dumas et al. 2011). The scale-dependent correlations (using wavelets) between the radio synchrotron and IR emissions are strong at large spatial scales but break down below a scale of a few 100 pc, which can be regarded as a measure of the electron diffusion length. The wavelet correlation between radio thermal and IR emission is strong at all scales because both quantities trace star formation.

Synchrotron intensity depends on the density of cosmic-ray electrons, which are accelerated in supernova remnants and diffuse into the interstellar medium, and on about the square of the strength of the total magnetic field B_{\perp} (► Sect. 2.2). Infrared intensity between about 20 and 70 μm (emitted from warm dust particles in thermal equilibrium, heated mainly by UV photons) is a measure of the star-formation rate. (Below about 20 μm large PAH particles and stars contribute; emission beyond about 70 μm comes from cold dust which is heated by the general radiation field.) Hence, the radio-infrared correlation can be presented as a correlation between turbulent field strength and star-formation rate (► Fig. 13-21). In contrast, the ordered field is either uncorrelated with the star-formation rate or anticorrelated in galaxies with magnetic arms for which the ordered field is strongest in interarm regions where star formation is low (► Sect. 4.4).

The radio-IR correlation requires that magnetic fields and star-formation processes are connected. In the “calorimeter” model, valid for starburst galaxies with strong fields where energy losses of the cosmic-ray electrons are strong, B^2 is assumed to increase with the infrared luminosity to obtain a linear radio-FIR correlation (Lisenfeld et al. 1996). In galaxies with low or medium star-formation rate (SFR), the cosmic-ray electrons can leave the galaxy and a combination of several processes with self-regulation is needed to explain the correlation within galaxies. If the dust is warm and optically thick to UV radiation, the IR intensity is proportional to the local SFR. Then, a possible scenario is the coupling of magnetic fields to the gas clouds



■ Fig. 13-21

Spiral galaxy NGC4254. Correlation between the strength of the total equipartition field (dominated by the turbulent field) and star-formation rate per area (determined from the 24 μm infrared intensities) within the galaxy, plotted on logarithmic scales. The slope of the fitted line gives an exponent of 0.18 ± 0.01 (Chyży 2008) (reproduced with permission © ESO)

($B \sim \rho^a$, where ρ is the neutral gas density), the Schmidt–Kennicutt law of star formation ($\text{SFR} \sim \rho^b$) (Niklas and Beck 1997). Depending on the values of the exponents a and b and whether or not equipartition between the energy densities of magnetic fields and cosmic rays is valid, a linear or nonlinear radio-IR correlation is obtained.

The radio-IR correlation also holds between the integrated luminosities of galaxies, which is one of the tightest correlations known in astronomy. Its explanation involves many physical parameters. The tightness needs multiple feedback mechanisms which are not yet understood (Lacki et al. 2010). The correlation holds for starburst galaxies up to redshifts of at least 4, although the average IR/radio ratio becomes smaller toward high redshifts (Murphy 2009). Nevertheless, the detection of radio emission in distant galaxies (which is at least partly of synchrotron origin) demonstrates that magnetic fields existed already in the early Universe.

Future radio telescopes like the SKA will allow the investigation of magnetic fields in young galaxies and search for their first fields (🔗 Sect. 5). Faraday rotation of polarized QSO emission in intervening galaxies also reveals magnetic fields in distant galaxies if they are regular on the spatial scale corresponding to the angular size of the background source. With this method, significant regular fields of several μG strengths on scales of about 10 kpc were discovered in galaxies up to redshifts of 2 (Kronberg et al. 2008). Detection of regular fields in young galaxies is a critical test of the mean-field dynamo theory (🔗 Sect. 5).

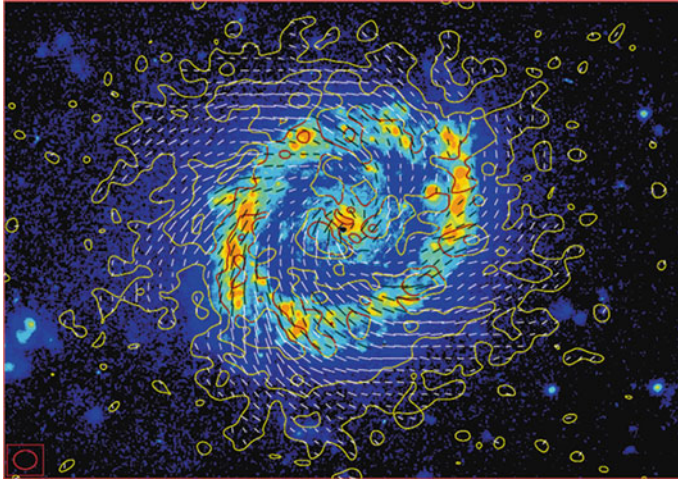
4.4 Magnetic Field Structures in Spiral Galaxies

4.4.1 Ordered Fields

At wavelengths ≤ 6 cm, Faraday rotation of the polarized synchrotron emission is generally small (except in central regions) so that the B-vectors directly trace the orientations of the ordered field (which can be regular or anisotropic, see 🔗 Sect. 2.3). Spiral patterns were found in almost every galaxy, even in those lacking optical spiral structure like the ringed galaxy NGC4736 (🔗 Fig. 13-22) and flocculent galaxies, while irregular galaxies show at most some patches of spiral structure (🔗 Sects. 4.6 and A.2). Spiral fields are also observed in the nuclear starburst regions of barred galaxies (🔗 Sect. 4.5). Galaxies of type Sa and S0 and elliptical galaxies without an active nucleus have little star formation and hence produce only few cosmic rays that could emit synchrotron emission. The only deep observation of a Sa galaxy, M104 with a prominent dust ring, revealed weak, ordered magnetic fields (Krause et al. 2006).

The gas flow in “smooth” galaxies (no bar, no tidal interaction, no strong density wave) is almost circular, while the field lines are spiral and do *not* follow the gas flow. If large-scale magnetic fields were frozen into the gas, differential rotation would wind them up to very small pitch angles. The observed smooth spiral patterns with significant pitch angles ($10\text{--}40^\circ$) indicate a general *decoupling* between magnetic fields and the gas flow due to magnetic diffusivity, which is a strong indication for mean-field dynamo action (🔗 Sect. 2.6). There is no other model to explain the magnetic spiral patterns in many types of galaxies.

However, the spiral pattern of magnetic fields cannot be solely the result of mean-field dynamo action. In gas-rich galaxies with strong density waves, the magnetic spiral pattern generally follows the spiral pattern of the gas arms. In the prototypical density-wave galaxy M51, for example, the pitch angle of the magnetic lines is mostly similar to that of the cold gas in the inner galaxy, but deviations occur in the outer parts of the galaxy, where the tidal effects of the



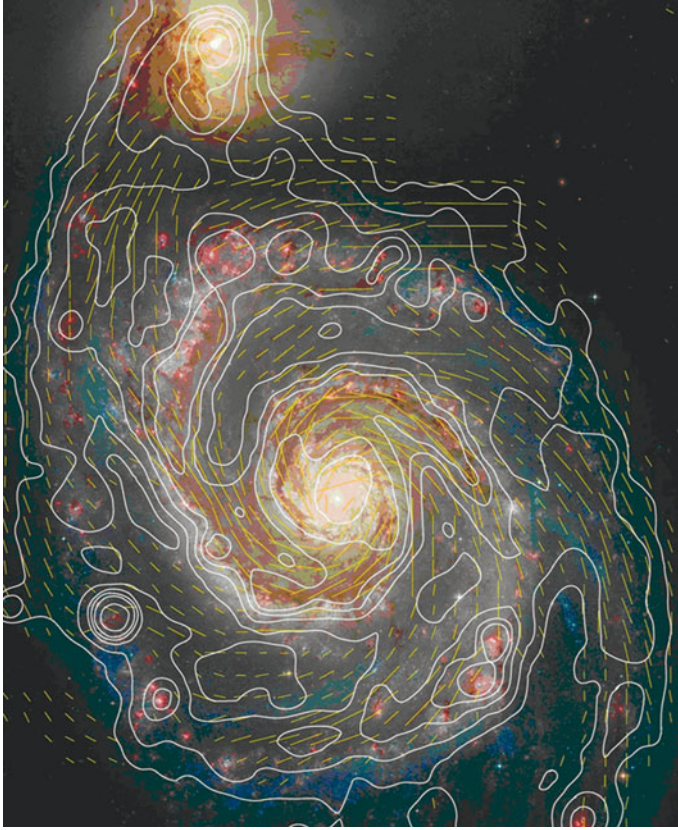
■ Fig. 13-22

Ring galaxy NGC4736. Polarized radio intensity (*contours*) and B-vectors at 3.6 cm, observed with the VLA (Chyży and Buta 2008). The background H α image is from Johan Hendrik Knapen (Inst. Astr. de Canarias) (reproduced by permission of the AAS)

companion galaxy are strong (Patrikeev et al. 2006). In dynamo theory, the pitch angle of the magnetic lines depends on global parameters (► Sect. 2.6) and is difficult to adjust to the pitch angle of the spiral structure of the gas. Furthermore, if the beautiful spiral pattern of M51 seen in radio polarization (► Fig. 13-23) is due to a regular field, it should be accompanied by a large-scale pattern in Faraday rotation, which is not observed. This means that most of the ordered field is *anisotropic* and probably generated by compression and shear of the non-axisymmetric gas flow in the density-wave potential. The anisotropic field is strongest at the positions of the prominent dust lanes on the inner edge of the inner gas spiral arms due to compression of turbulent fields in the density-wave shock. Anisotropic fields also fill the interarm space, without signs of compression, probably generated by shearing flows. Regular fields also exist but are much weaker (see below). In the outer galaxy, ordered fields coincide with the outer southern and southwestern spiral arms; these are possibly tidal arms with strong shear. The northeastern field deviates from the gas arm and points toward the companion, signature of the interaction.

M83 (► Fig. 13-24) and NGC2997 (Han et al. 1999) are cases similar to M51, with enhanced ordered (anisotropic) fields at the inner edges of the inner optical arms, ordered fields in interarm regions, and ordered fields coinciding with the outer optical arms. Density-wave galaxies with less star-formation activity, like M81 (Krause et al. 1989b) and NGC1566 (Ehle et al. 1996), show little signs of field compression, and the ordered fields occur mainly in the interarm regions.

Observations of another gas-rich spiral galaxy, NGC6946, revealed a surprisingly regular distribution of polarized emission with two symmetric *magnetic arms* located in interarm regions, with orientations parallel to the adjacent optical spiral arms and no signs of compression at the inner edge of the gas arms (► Fig. 13-25). Their degree of polarization is exceptionally high (up to 50%); the field is almost totally ordered and mostly regular, as indicated by Faraday rotation measures. With the higher sensitivity at 20 cm wavelength, more magnetic arms



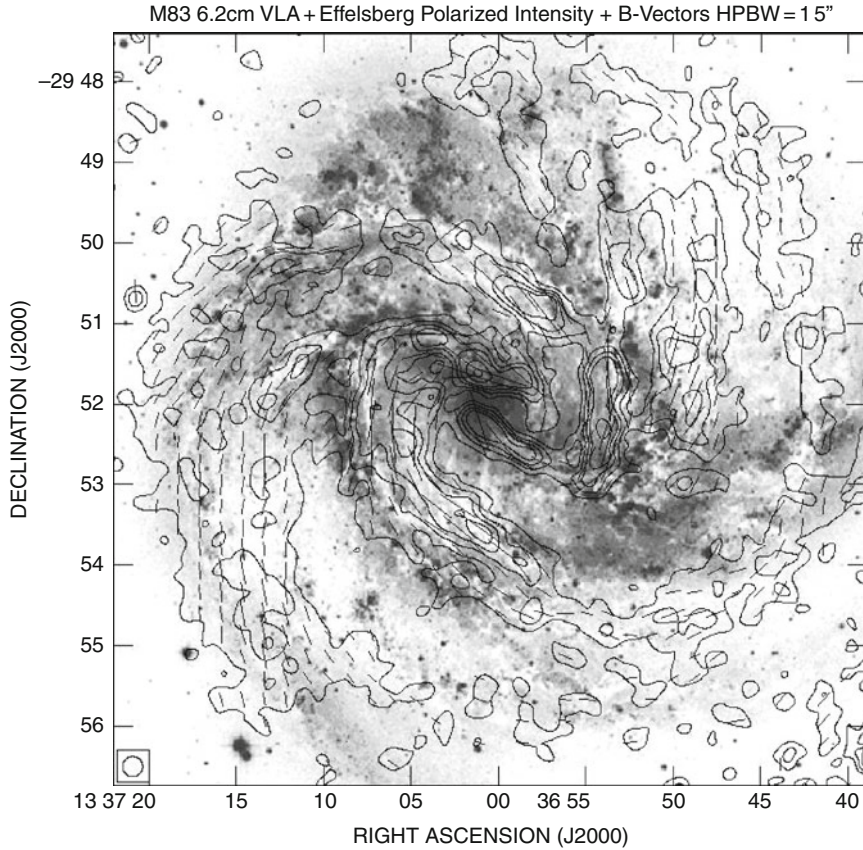
■ Fig. 13-23

Spiral galaxy M51. Total radio intensity (*contours*) and B-vectors at 6 cm wavelength, combined from observations with the VLA and Effelsberg 100-m telescopes (Fletcher et al. 2011). The background optical image is from the HST (Hubble Heritage Team) (Graphics: Sterne und Weltraum)

appear in the northern half of NGC6946, extending far beyond the optical arms, but located between outer HI arms. Magnetic arms have also been found in M83 (► Fig. 13-24), NGC2997, and several other gas-rich spiral galaxies.

Ordered magnetic fields may also form spiral features that are disconnected from the optical spiral pattern. Long, highly polarized filaments were discovered in the outer regions of IC342, where only faint arms of HI line emission exist (Krause et al. 1989a). More recent observations at 20 cm revealed a system of such features extending to large distances from the center (► Fig. 13-27).

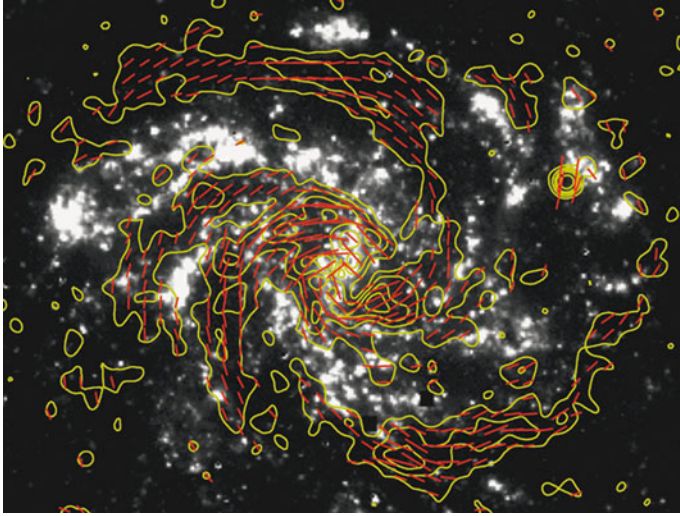
In the highly inclined Andromeda galaxy, M31, (► Fig. 13-26) the spiral arms are hard to distinguish due to the insufficient angular resolution. Star-formation activity is concentrated to a limited radial range at around 10 kpc distance from the center (the “ring”). The ordered fields are strongest in the massive dust lanes where the degree of polarization is about 40%. The field follows the “ring” with a coherent direction (► Fig. 13-29).



■ Fig. 13-24

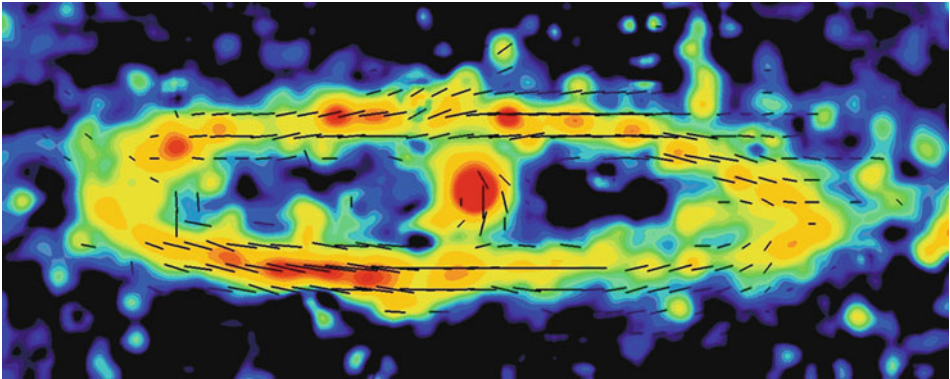
Barred galaxy M83. Polarized radio intensity (*contours*) and B-vectors at 6 cm, combined from observations with the VLA and Effelsberg telescopes (Beck et al. unpublished). The background optical image is from Dave Malin (Anglo Australian Observatory)

At wavelengths of around 20 cm, a striking asymmetry of the polarized emission occurs along the major axis of all 12 spiral galaxies observed so far with sufficiently high sensitivity that have inclinations of less than about 60° . The emission is almost completely depolarized by Faraday dispersion, e.g., in IC342 (● Fig. 13-27) on one side of the major axis, which is always the kinematically *receding* one (positive radial velocities). In strongly inclined galaxies, both sides of the major axis become Faraday-depolarized at around 20 cm, as a result of the long pathlength. The asymmetry is still visible at 11 cm but disappears at smaller wavelengths. This tells us that, in addition to spiral fields in the disk, fields in the halo are needed, as predicted by mean-field dynamo models (Braun et al. 2010; Urbanik et al. 1997; see ● Sect. 4.7). The effect of such halo fields becomes prominent at 20 cm because most of the polarized emission from the disk is Faraday-depolarized (● Sect. 2.4). Testing by observations at longer wavelengths will soon become possible with LOFAR (● Sect. 5).



■ Fig. 13-25

Spiral galaxy NGC6946. Polarized radio intensity (*contours*) and B-vectors at 6 cm, combined from observations with the VLA and Effelsberg 100-m telescopes (Beck 2007). The background H α image is from Anne Ferguson (Graphics: Sterne und Weltraum)

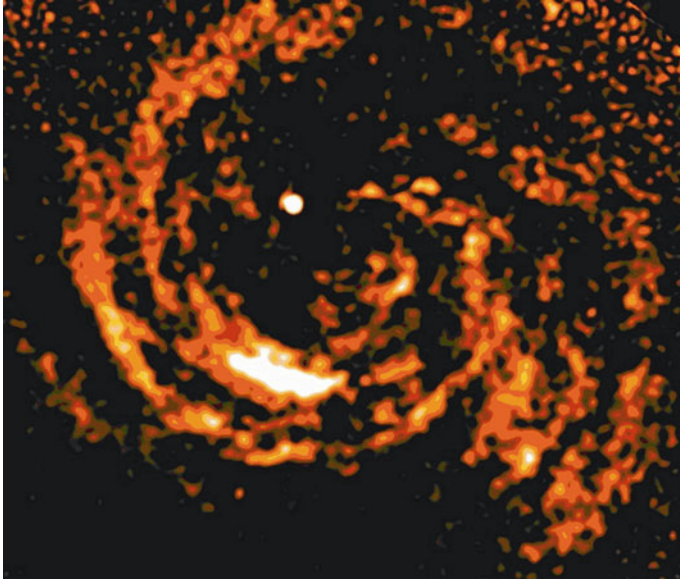


■ Fig. 13-26

Spiral galaxy M31. Total radio intensity (*colors*) and B-vectors (corrected for Faraday rotation) at 6 cm, observed with the Effelsberg telescope (Berkhuijsen et al. 2003)

4.4.2 Regular Fields

Ordered magnetic fields as observed by polarized emission can be anisotropic (see above) or regular (with a coherent direction). *Faraday rotation measures* (RM) are signatures of such regular fields. RM is determined from multiwavelength radio polarization observations (● Sect. 2.4). Spiral dynamo modes (● Sect. 2.6) can be identified from the periodicity of the



■ Fig. 13-27

Spiral galaxy IC342. Polarized radio intensity at 20 cm, combined from VLA C- and D-array observations. The field size is about $30' \times 28'$. The inner and northwestern parts are depolarized at this wavelength (Beck, unpublished)

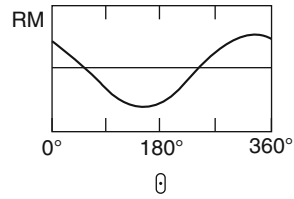
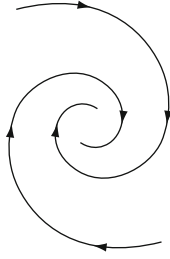
azimuthal variation of RM in inclined galaxy disks (► Fig. 13-28), where the RM can be determined from diffuse polarized emission (Krause 1990) or from RM data of polarized background sources (Stepanov et al. 2008). If several dynamo modes are superimposed, Fourier analysis of the RM variation is needed. The resolution of present-day observations is sufficient to identify not more than 2–3 modes.

The disks of a few spiral galaxies indeed reveal large-scale RM patterns giving strong evidence for modes generated by the mean-field dynamo. M31 is the prototype of a dynamo-generated magnetic field (● Fig. 13-29). The discovery became possible thanks to the large angular extent and the high inclination of M31. The polarized intensity at 6 cm is largest near the minor axis where the field component B_{\perp} is largest (► Fig. 13-30a), while the maxima in $|RM|$ are observed near the major axis where the line-of-sight field component B_{\parallel} is strongest (► Fig. 13-30b). This single-periodic RM variation is a clear signature of a dominating axisymmetric spiral (ASS) disk field (dynamo mode $m=0$) (Fletcher et al. 2004), which extends to at least 25 kpc distance from the center when observed with an *RM grid* (see below) (Han et al. 1998).

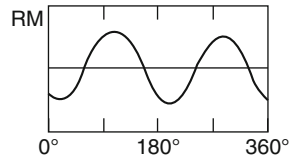
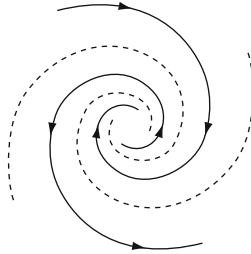
Other galaxies with a dominating axisymmetric disk field are the nearby spiral IC342, the Virgo galaxy NGC4254, the almost edge-on galaxies NGC253, NGC891, and NGC5775, the irregular Large Magellanic Cloud (LMC), and a few further candidates (see Appendix).

By measuring the signs of the RM distribution and the velocity field on both sides of a galaxy's major axis, the *inward* and *outward* directions of the radial component of the ASS field can be easily distinguished (► Fig. 13-31). Dynamo models predict that both signs have the same probability, which is confirmed by observations. The ASS fields of M31, IC342, NGC253,

Axisymmetric Spiral Structure

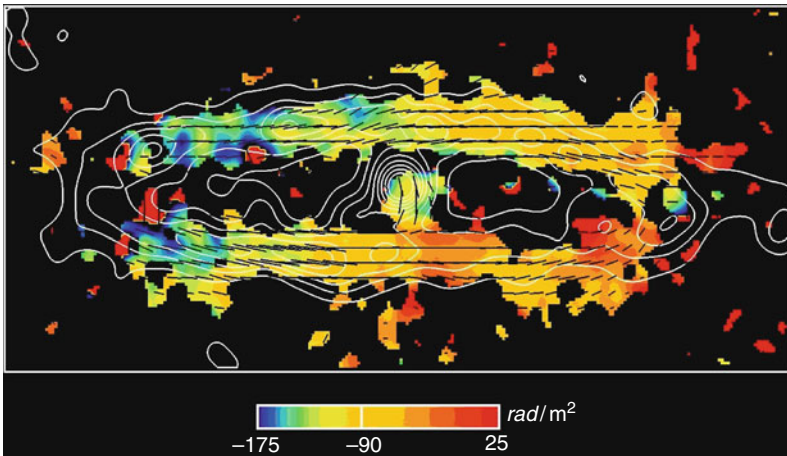


Bisymmetric Spiral Structure



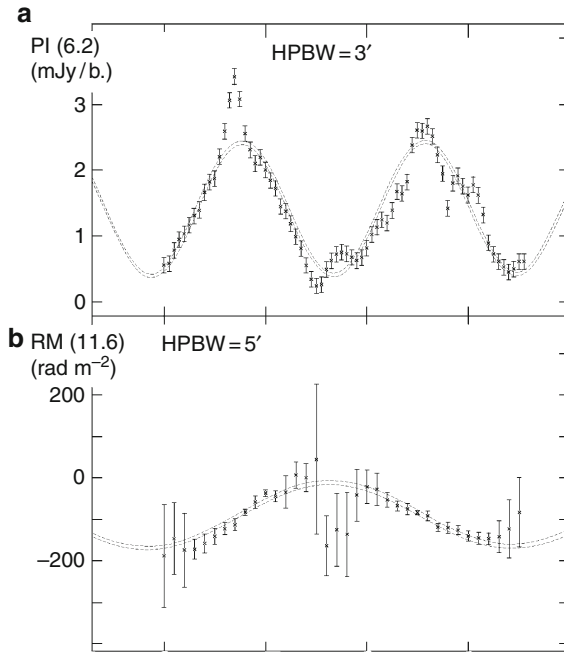
■ Fig. 13-28

Azimuthal RM variations (measured from the major axis) for axisymmetric spiral (ASS) and bisymmetric spiral (BSS) fields in inclined galaxies (Krause 1990)



■ Fig. 13-29

Spiral galaxy M31. Total radio intensity at 6 cm (*contours*), B-vectors and Faraday rotation measures between 6 and 11 cm (*colors*), derived from observations with the Effelsberg telescope (Berkhuijsen et al. 2003). The average rotation measure of about -90 rad m^{-2} is caused by the foreground medium in the Milky Way



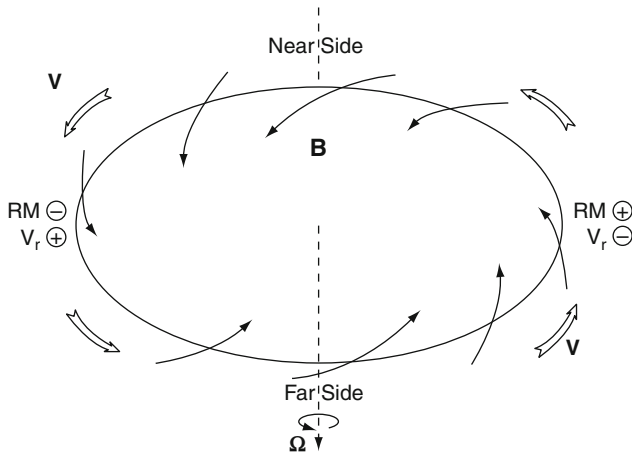
■ Fig. 13-30

Spiral galaxy M31. (a) Variation of polarized intensity and (b) Faraday rotation measures between 6 and 11 cm along the azimuthal angle in the plane of the galaxy, counted counterclockwise from the northern major axis (left side in [Fig. 13-29](#)) (Berkhuijsen et al. 2003) (reproduced with permission © ESO)

and the ASS field component in NGC6946 point inward, while those of NGC4254, NGC5775, and the ASS component of the disk field in M51 point outward.

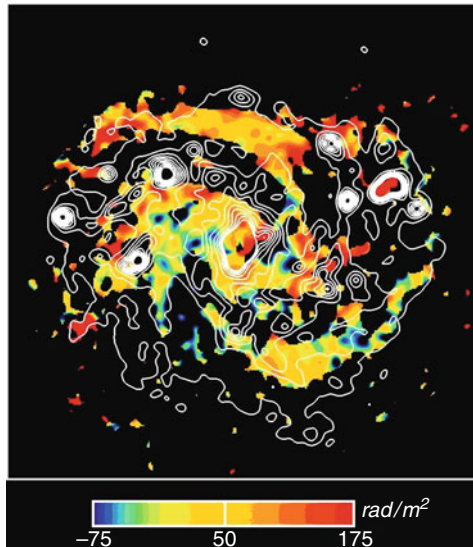
M81, M83, and an intervening galaxy at a redshift of 0.4 in front of the quasar PKS1229-021 are the only candidates so far for a bisymmetric spiral (BSS) field ($m = 1$), characterized by a double-periodic RM variation, but the data quality is limited in all these cases. Dominating BSS fields are rare, as predicted by dynamo models. It was proposed that tidal interaction can excite the BSS mode, but no preference for BSS was found even in the most heavily interacting galaxies in the Virgo cluster ([Sect. 4.8](#)). The idea that galactic fields are wound-up primordial intergalactic fields that are of BSS type ([Sect. 2.6](#)) can also be excluded from the existing observations.

Faraday rotation in NGC6946 and in other similar galaxies with magnetic arms can be described by a superposition of two azimuthal dynamo modes ($m = 0$ and 2) with about equal amplitudes, where the quadrisymmetric spiral (QSS) $m = 2$ mode is phase shifted with respect to the density wave (Beck 2007). This model is based on the RM pattern of NGC6946 that shows different field directions in the northern and southern magnetic arm ([Fig. 13-32](#)). A weaker QSS mode superimposed onto the dominating ASS mode is indicated in the disk of M51 and in the inner part of M31. A superposition of ASS and BSS modes can describe the fields of M33



■ Fig. 13-31

The sign of the Faraday rotation measure RM and the sign of the rotation velocity component v_r along the line of sight, measured near the major axis of a galaxy, are opposite in the case of the inward direction of the radial component of an ASS-type field, while the signs are the same for the outward field direction. Trailing spirals are assumed (Krause and Beck 1998) (reproduced with permission © ESO)



■ Fig. 13-32

Spiral galaxy NGC6946. Total radio intensity at 6 cm (*contours*) and Faraday rotation measures between 3.6 and 6 cm (*colors*), derived from combined observations with the VLA and Effelsberg telescopes (Beck 2007). The average rotation measure of about $+50 \text{ rad m}^{-2}$ is caused by the foreground medium in the Milky Way

and NGC4254, while three modes (ASS+BSS+QSS) are needed for NGC1097, NGC1365, and NGC4414 (see Appendix).

In most galaxies observed so far, a spiral polarization pattern was found, but no large-scale RM pattern as a signature of regular fields. In many cases, the available polarization data is insufficient to derive reliable RMs. In other cases, the data quality is high but no large-scale RM patterns are visible. In density-wave galaxies, strong compression and shearing flows generate *anisotropic* fields (with frequent reversals) of spiral shape which are much stronger than the underlying regular field, like in M51 (see above). In galaxies without density waves, several dynamo modes may be superimposed but cannot be distinguished with the limited sensitivity and resolution of present-day telescopes. Another explanation is that the timescale for the generation of large-scale modes is longer than the galaxy's lifetime so that the regular field is not fully organized and still restricted to small regions.

Large-scale *field reversals* were discovered from pulsar RMs in the Milky Way (▶ Sect. 3.5), but nothing similar has yet been detected in spiral galaxies, although high-resolution RM maps of Faraday rotation are available for many spiral galaxies. In M81, the dominating BSS field implies two large-scale reversals (Krause et al. 1989b). The disk fields of several galaxies can be described by a mixture of modes where reversals may emerge in a limited radial and azimuthal range of the disk, like in NGC4414 (Soida et al. 2002). However, no multiple reversals along the radial direction, like those in the Milky Way, were found so far in the disk of any external galaxy. A satisfying explanation is still lacking (▶ Sect. 3.5). Reversals on smaller scales are probably frequent but difficult to observe in external galaxies with the resolution of present-day telescopes. Only in the barred galaxy NGC7479, where a jet serves as a bright polarized background (▶ Fig. 13-50), several reversals on 1–2 kpc scale were detected in the foreground disk of the galaxy (Laine and Beck 2008).

While the azimuthal symmetry of the dynamo modes is known for many galaxies, the vertical symmetry (*even* or *odd*) is much harder to determine. The RM patterns of even and odd modes are similar in mildly inclined galaxies. The toroidal field of odd modes reverses its sign above and below the galactic plane. Thus, in a mildly inclined odd field, half of the RM is observed compared to that in an even field, which cannot be distinguished in view of the large RM variations caused by ionized gas density and field strength. The symmetry type becomes only visible in strongly inclined galaxies as a RM sign reversal above and below the plane. Only even fields were found so far (in M31, NGC253, NGC891, and NGC5775), which is again in agreement with the prediction of dynamo models (▶ Sect. 4.7).

If polarized emission is too weak to be detected, the method of *RM grids* toward polarized background QSOs can still be applied. This allows the determination of a large-scale field pattern in an intervening galaxy on the line of sight (Kronberg et al. 1992). Here, the distance limit is given by the polarized flux of the background QSO which can be much larger than that of the intervening galaxy so that this method can be applied to much larger distances than the analysis of RM of the polarized emission from the foreground galaxy itself. At least 10 randomly distributed background sources behind the galaxy disk are needed to recognize simple patterns and several 1,000 sources for a full reconstruction (Stepanov et al. 2008). Present-day observations are not sensitive enough, and one has to wait for the SKA and its precursor telescopes.

Ordered fields of nearby galaxies seen edge-on near the disk plane are preferably oriented parallel to the plane (▶ Sect. 4.7). As a result, polarized emission can be detected from distant, *unresolved* galaxies if they are symmetric (not distorted by interaction) and their inclination is larger than about 20° (Stil et al. 2009). This opens another method to search for ordered fields in distant galaxies. As the plane of polarization is almost independent of wavelength, distant spiral

galaxies with known orientation of their major axis can also serve as background polarized sources to search for Faraday rotation by intergalactic fields in the foreground.

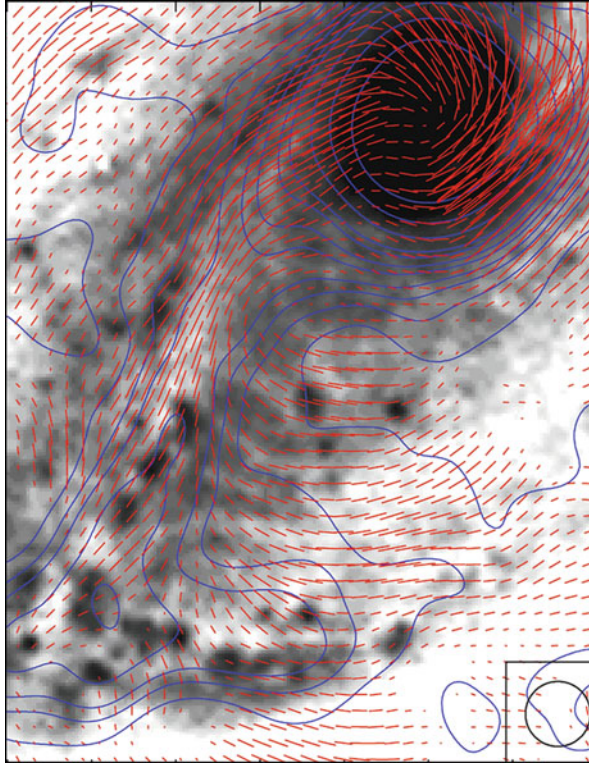
In summary, magnetic fields in spiral galaxies are complex. The observations can best be explained as a superposition of dynamo-generated modes of regular fields coupled to the diffuse warm gas, plus anisotropic fields by shearing and compressing flows, plus turbulent fields coupled to the cold gas. The magnetic fields in barred galaxies behave similarly (➤ Sect. 4.5). For a more detailed model of the physics of the field-gas interaction, high-resolution data with future telescopes are required (➤ Sect. 5).

4.5 Magnetic Fields in Barred Galaxies

Gas and stars in the gravitational potential of strongly barred galaxies move in highly noncircular orbits. Numerical models show that gas streamlines are deflected in the bar region along shock fronts, behind which the cold gas is compressed in a fast shearing flow (Athanasoula 1992). The compression regions traced by massive dust lanes develop along the edge of the bar that is leading with respect to the galaxy's rotation because the gas rotates faster than the bar pattern. The warm, diffuse gas has a higher sound speed and is not compressed. According to simulations, the shearing flows around a bar should amplify magnetic fields and generate complicated field patterns changing with time (Otmianowska-Mazur et al. 2002). The asymmetric gas flow may also enhance dynamo action and excite the QSS ($m = 2$) mode (Moss et al. 2001).

Twenty galaxies with large bars were observed with the Very Large Array (VLA) and with the Australia Telescope Compact Array (ATCA) (Beck et al. 2002, 2005a). The total radio luminosity (a measure of the total magnetic field strength) is strongest in galaxies with high far-infrared luminosity (indicating high star-formation activity), a result similar to that in non-barred galaxies. The average radio intensity, radio luminosity, and star-formation activity all correlate with the relative bar length. Polarized emission was detected in 17 of the 20 barred galaxies. The pattern of the regular field in the galaxies with long bars (NGC1097, 1365, 1559, 1672, 2442, and 7552) is significantly different from that in non-barred galaxies: Field enhancements occur outside of the bar (upstream), and the field lines are oriented at large angles with respect to the bar.

NGC1097 (➤ Fig. 13-33) is one of the nearest barred galaxies and hosts a huge bar of about 16 kpc length. The total radio intensity (not shown in the figure) and the polarized intensity are strongest in the downstream region of the dust lanes (southeast of the center). This can be explained by a compression of turbulent fields in the bar's shock, leading to strong and anisotropic fields in the downstream region. The surprising result is that the polarized intensity is also strong in the upstream region (south of the center in ➤ Fig. 13-33), where RM data indicate that the field is regular. The pattern of field lines in NGC1097 is similar to that of the gas streamlines as obtained in numerical simulations (Athanasoula 1992). This suggests that the ordered (partly regular) magnetic field is aligned with the flow and amplified by strong shear. Remarkably, the optical image of NGC1097 shows dust filaments in the upstream region which are almost perpendicular to the bar and thus aligned with the ordered field. Between the region upstream of the southern bar and the downstream region, the field lines smoothly change their orientation by almost 90° . The ordered field is probably coupled to the diffuse gas and thus avoids being shocked in the bar. The magnetic energy density in the upstream region is sufficiently high to affect the flow of the diffuse gas.

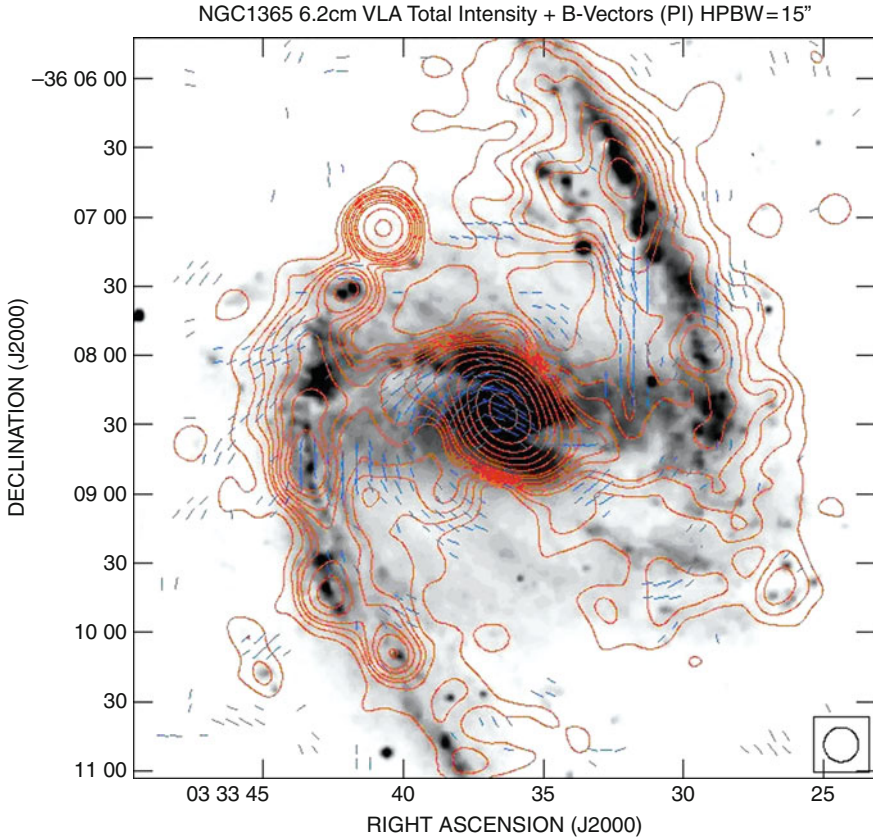


■ Fig. 13-33

Southern half of the barred galaxy NCC1097. Total radio intensity (*contours*) and B-vectors at 3.5 cm, observed with the VLA (Beck et al. 2005a). The background optical image is from Halton Arp (Cerro Tololo Observatory)

NGC1365 (► Fig. 13-34) is similar to NGC1097 in its overall properties, but the polarization data indicate that the shear is weaker. The ordered field bends more smoothly from the upstream region into the bar, again with no indication of a shock. M83 is the nearest barred galaxy but with a short bar; it shows compressed ordered fields at the leading edges of the bar on both sides of the nucleus and some polarization in the upstream regions (► Fig. 13-24). In all other galaxies observed so far (Sect. A.2), the resolution is insufficient to separate the bar and upstream regions.

The central regions of barred galaxies are often sites of ongoing intense star formation and strong magnetic fields that can affect the gas flow. Radio emission from ring-like regions has been found in NGC1097, NGC1672, and NGC7552 (Beck et al. 2005b). NGC1097 hosts a bright ring with about 1.5 kpc diameter and an active nucleus in its center (► Fig. 13-35). The ordered field in the ring has a spiral pattern and extends toward the nucleus. The orientation of the innermost spiral field agrees with that of the spiral dust filaments visible on optical images. Magnetic stress in the circumnuclear ring can drive mass inflow at a rate of $dM/dt = -h/\Omega$ ($< b_r b > + B_r B_\Phi$), where h is the scale height of the gas, Ω its angular rotation velocity, b the strength of the turbulent field and B that of the ordered field, and r and Φ denote the radial



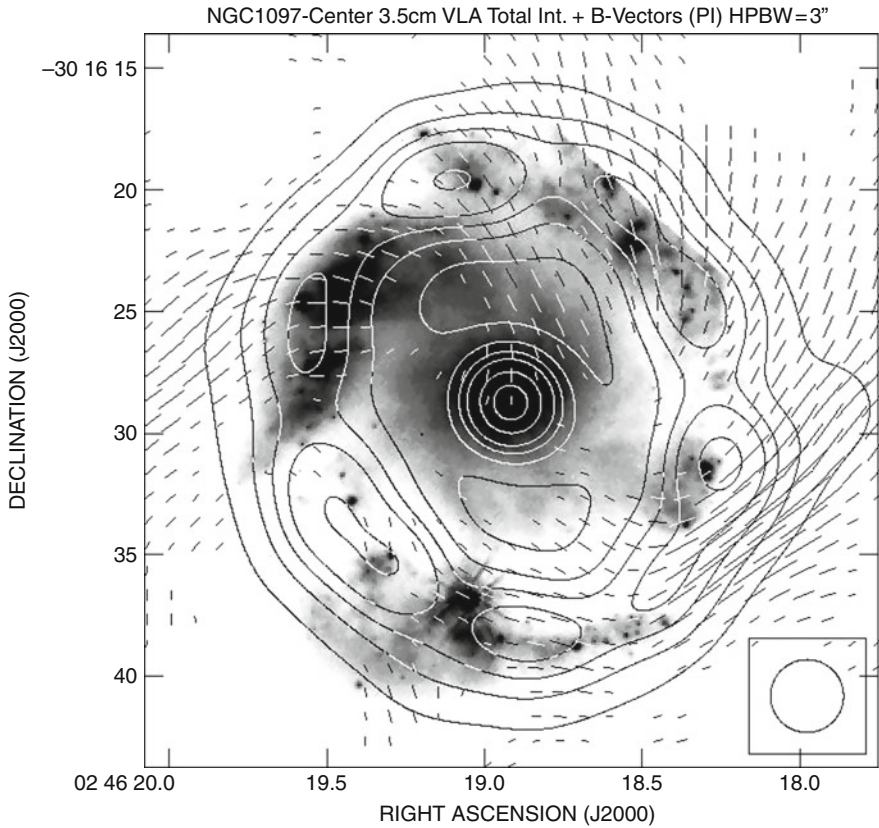
■ Fig. 13-34

Barred galaxy NGC1365. Total radio intensity (*contours*) and B-vector at 6 cm, observed with the VLA (Beck et al. 2005a). The background optical image is from Per Olof Lindblad (ESO)

and azimuthal field components (Balbus and Hawley 1998). For NGC1097, $h \approx 100$ pc, $v \approx 450$ km s⁻¹ at 1 kpc radius, $b_r \approx b_\phi \approx 50$ μ G gives an inflow rate of several M_\odot /year, which is sufficient to fuel the activity of the nucleus (Beck et al. 2005a).

In summary, the turbulent field in galaxies with massive bars is coupled to the cold gas and compressed in the bar's shock. The ordered field outside the bar region follows the general flow of the cold and warm gas, possibly due to shear, but decouples from the cold gas in front of the shock and goes with the diffuse warm gas. The polarization pattern in barred galaxies can be used a tracer of the flow of diffuse gas in the sky plane and hence complements spectroscopic measurements of radial velocities. Detailed comparisons between polarimetric and spectroscopic data are required, as well as MHD models, including the back-reaction of the magnetic fields onto the gas flow.

Radio polarization data have revealed differences but also similarities between the behaviours of regular magnetic fields in barred and non-barred galaxies. In galaxies without bars and without strong density waves, the field lines have a spiral shape, they do not follow the gas flow and are probably amplified by dynamo action. In galaxies with massive bars or strong



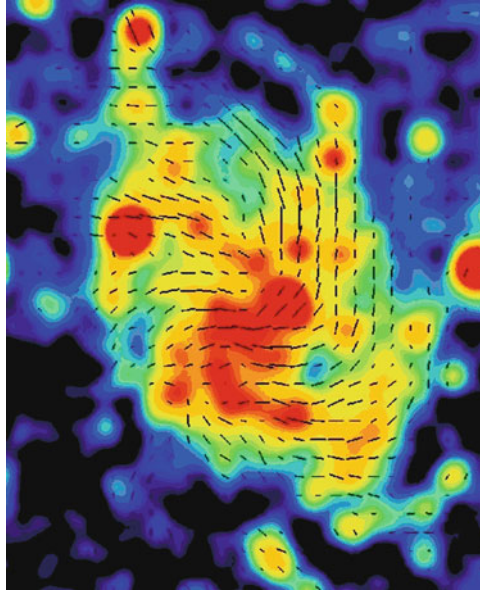
■ Fig. 13-35

Central star-forming ring of the barred galaxy NGC1097. Total radio intensity (*contours*) and B-vectors at 3.5 cm, observed with the VLA (Beck et al. 2005a). The background optical image is from the Hubble Space Telescope (reproduced with permission © ESO)

density waves, the field lines mostly follow the flow of the diffuse warm gas. Near the shock fronts galaxies with strong bars and with strong density waves (● Sect. 4.4) reveal a similar behaviour: Turbulent fields are coupled to the cold gas, are shocked, and become anisotropic, while regular fields are coupled to the warm diffuse gas and hence avoid the shock.

4.6 Flocculent and Irregular Galaxies

Flocculent galaxies have disks but no prominent spiral arms. Nevertheless, spiral magnetic patterns are observed in all flocculent galaxies, indicative that the mean-field dynamo works independently of density waves. The multiwavelength data of M33 and NGC4414 call for a mixture of dynamo modes or an even more complicated field structure (Sect. A.2). Ordered magnetic fields with strengths similar to those in grand-design spiral galaxies have been detected in the flocculent galaxies M33 (● Fig. 13-36), NGC3521, NGC5055, and in NGC4414,



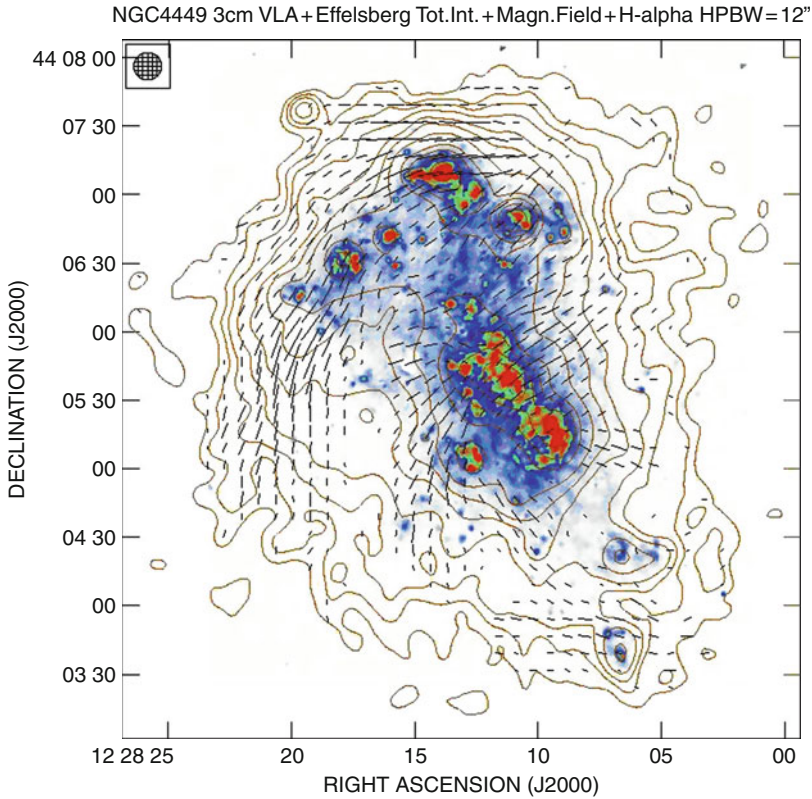
■ Fig. 13-36

Flocculent galaxy M33. Total radio intensity (colors) and B-vectors of the flocculent galaxy at 3.6 cm, observed with the Effelsberg telescope (Tabatabaei et al. 2008)

and also the mean degree of polarization (corrected for the differences in spatial resolution) is similar between grand-design and flocculent galaxies (Knapik et al. 2000).

Radio continuum maps of *irregular*, slowly rotating galaxies may reveal strong total equipartition magnetic fields, e.g., in the Magellanic-type galaxy NGC4449 (► Fig. 13-37) and in IC10 (► Fig. 13-38). In NGC4449, a fraction of the field is ordered with about $7 \mu\text{G}$ strength and a spiral pattern. Faraday rotation shows that this ordered field is partly regular and the mean-field dynamo is operating in this galaxy. The total field is of comparable strength ($10\text{--}15 \mu\text{G}$) in starburst dwarfs like NGC1569 (Kepley et al. 2010) where star-formation activity is sufficiently high for the operation of the small-scale dynamo (► Sect. 2.6). In these galaxies, the energy density of the magnetic fields is only slightly smaller than that of the (chaotic) rotation of the gas and thus may affect the evolution of the whole system. The starburst dwarf galaxy NGC1569 shows polarized emission, but no large-scale regular field. In dwarf galaxies with very weak star-forming activity, no polarized emission is detected and the turbulent field strength is several times smaller than in spiral galaxies (Chyży et al. 2011), sometimes less than $5 \mu\text{G}$ (Chyży et al. 2003). The latter value may indicate a sensitivity limit of present-day observations or a threshold for small-scale dynamo action.

The Magellanic Clouds are the closest irregular galaxies and deserve special attention. Polarization surveys with the Parkes single-dish telescope at several wavelengths had low angular resolution and revealed weak polarized emission. Two magnetic filaments were found in the LMC south of the 30 Dor star-formation complex (Klein et al. 1993). ATCA surveys of an RM grid toward background sources show that the LMC probably contains a large-scale magnetic field similar to large spirals (Gaensler et al. 2005) and that the SMC is weak and uniformly



■ Fig. 13-37

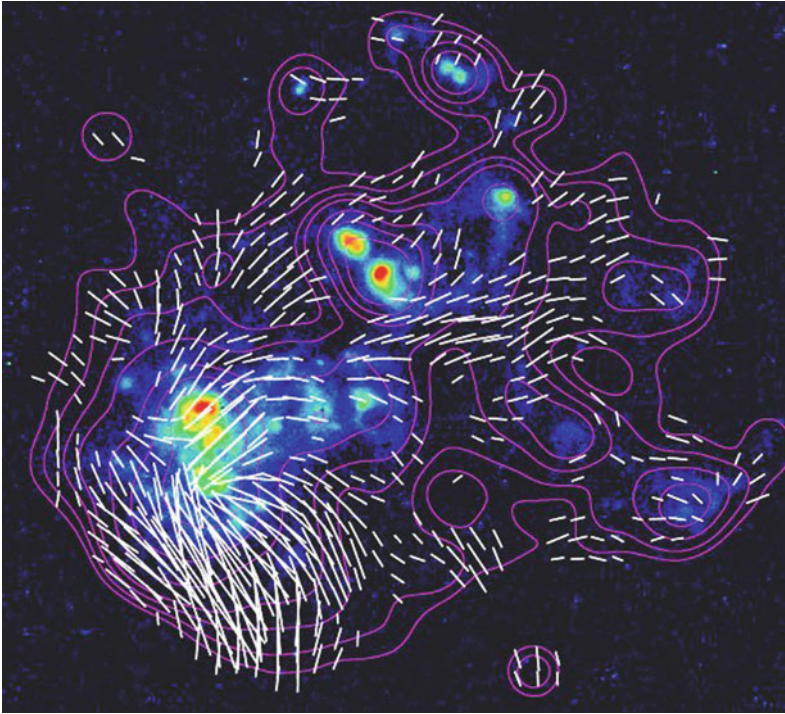
Magellanic-type galaxy NGC4449. Total radio intensity (*contours*) and B-vectors at 3.6 cm, combined from VLA and Effelsberg observations (Chyży et al. 2000). The background image shows the H α emission (reproduced with permission © ESO)

directed away from us, possibly part of a pan-Magellanic field joining the two galaxies (Mao et al. 2008).

4.7 Radio Halos

Radio halos are observed around the disks of most edge-on galaxies, but their radio intensity and extent varies significantly. The halo luminosity in the radio range correlates with those in H α and X-rays (Tüllmann et al. 2006), although the detailed halo shapes vary strongly between the different spectral ranges. These results suggest that star formation in the disk is the energy source for halo formation and the halo size is determined by the energy input from supernova explosions per surface area in the projected disk (Dahlem et al. 1995).

In spite of the different intensities and extents of radio halos, their exponential scale heights at 5 GHz are about 1.8 kpc (Dumke and Krause 1998; Heesen et al. 2009a), with a surprisingly

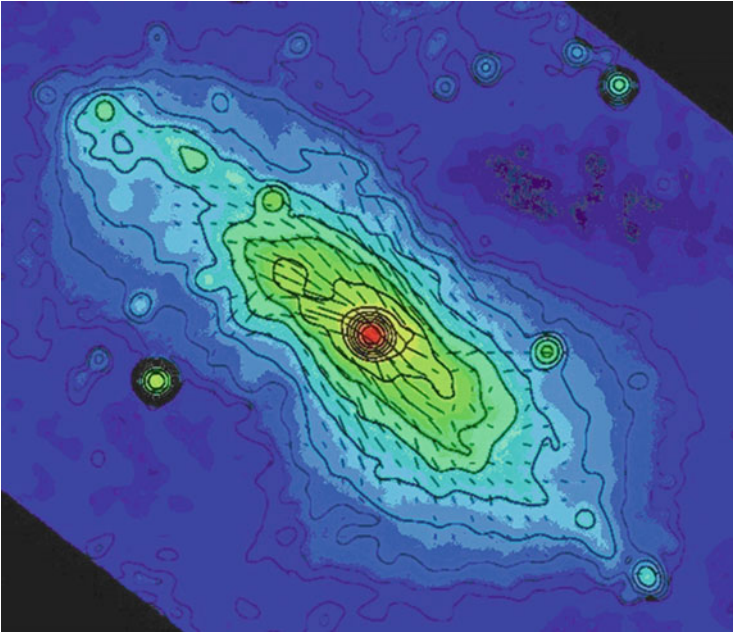


■ Fig. 13-38

Irregular galaxy IC10. Total radio intensity (*contours*) and B-vectors at 6 cm, observed with the VLA (Chris Chyży, Kraków University). The background H α image is from Dominik Bomans (Bochum University)

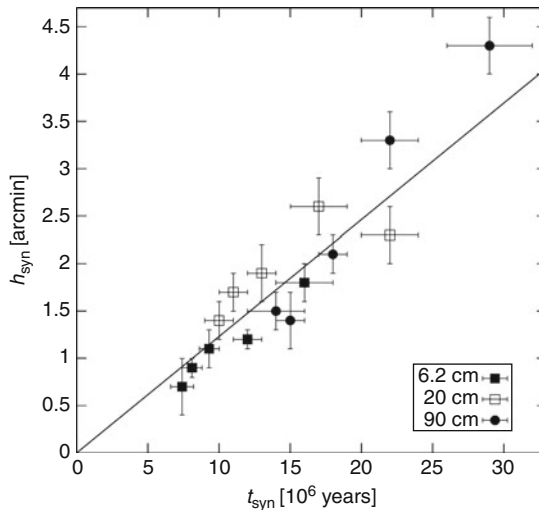
small scatter in the sample, ranging from one of the weakest halos, NGC4565, to the brightest ones known, NGC253 (► Fig. 13-39) and NGC891 (► Fig. 13-41). In case of equipartition between the energy densities of magnetic field and cosmic rays, the exponential scale height of the total field is at least $(3-\alpha)$ times larger than the synchrotron scale height (where $\alpha \approx -1$ is the synchrotron spectral index), ≥ 7 kpc. The real value depends on the energy losses of the cosmic-ray electrons propagating into the halo (► Sect. 2.2). A prominent exception is NGC4631 with the largest radio halo observed so far (► Fig. 13-42). With large-scale heights, the magnetic energy density in halos is much higher than that of the thermal gas, while still lower than that of the dominating kinetic energy of the gas outflow.

Radio halos grow in size with decreasing observation frequency. The extent is limited by energy losses of the cosmic-ray electrons, i.e., synchrotron, inverse Compton, and adiabatic losses (Heesen et al. 2009a). The stronger magnetic field in the central region causes stronger synchrotron loss, leading to the “dumbbell” shape of many radio halos, e.g., around NGC253 (► Fig. 13-39). From the radio scale heights of NGC253 at three frequencies and the electron lifetimes (due to synchrotron, inverse Compton, and adiabatic losses), an outflow bulk speed of about 300 km s^{-1} was measured (► Fig. 13-40). The similarity of the scale height of the radio halos around most edge-on galaxies observed so far, in spite of the different field strengths and



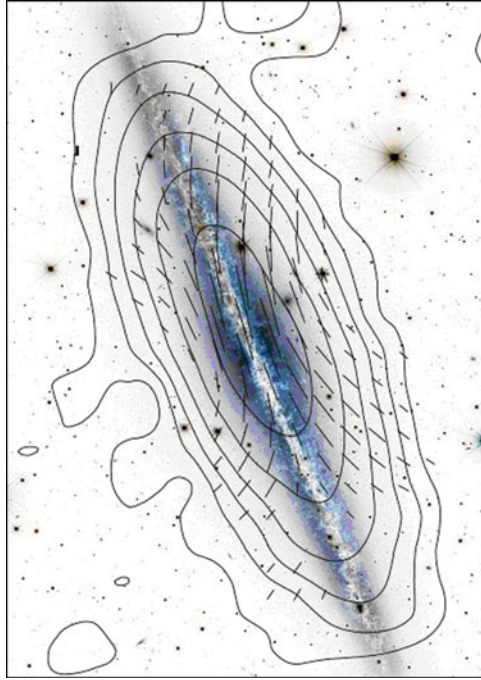
■ Fig. 13-39

Almost edge-on spiral galaxy NGC253. Total radio intensity (*contours*) and B-vectors at 6 cm, combined from observations with the VLA and the Effelsberg telescope (Heesen et al. 2009b)



■ Fig. 13-40

Synchrotron scaleheights of the northern radio halo of NGC253 at different distances from the center and at different wavelengths, as a function of synchrotron lifetime of cosmic-ray electrons. The slope of the linear fit corresponds to a bulk outflow speed of about 300 km s^{-1} (Heesen et al. 2009a) (reproduced with permission © ESO)



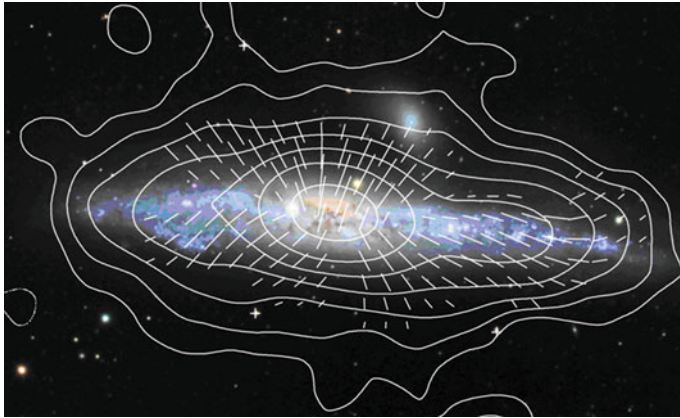
■ Fig. 13-41

Edge-on spiral galaxy NGC891. Total radio intensity (*contours*) and B-vectors at 3.6 cm wavelength, observed with the Effelsberg telescope (Krause 2009). The background optical image is from the CFHT

hence different electron lifetimes, indicates that the outflow speed increases with the average strength of the total field and with the star-formation rate (Krause 2009). Outflows slower than the escape velocity are often called *fountain flows*, while escaping flows are *galactic winds*.

Radio polarization observations of nearby galaxies seen edge-on generally show a disk-parallel field near the disk plane (Dumke et al. 1995). High-sensitivity observations of several edge-on galaxies like NGC253 (► Fig. 13-39), NGC891 (● Fig. 13-41), NGC5775 (Soida et al. unpubl.; Tüllmann et al. 2000), and M104 (Krause et al. 2006) revealed vertical field components which increase with increasing height above and below the galactic plane and also with increasing radius, the so-called *X-shaped* halo fields. The X-pattern is even seen in NGC4565 with its low star-formation rate and a radio-faint halo, thus this pattern seems to be a general phenomenon.

The observation of X-shaped field patterns is of fundamental importance to understand the field origin in halos. The field is probably transported from the disk into the halo by an outflow emerging from the disk. The X-shaped halo field is consistent with the predictions from mean-field dynamo models if outflows with moderate velocities are included (► Sect. 2.6). Numerical models (neglecting magnetic fields) indicate that global gas outflows from the disks of young galaxies can also be X-shaped due to pressure gradients (Dalla Vecchia and Schaye 2008). MHD models are still lacking.



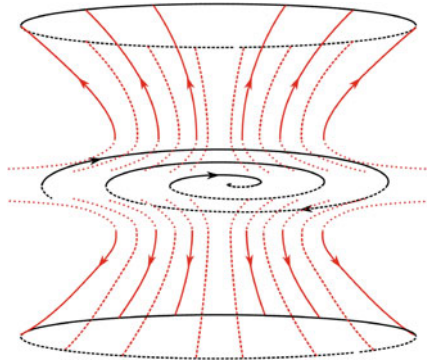
■ Fig. 13-42

Edge-on irregular galaxy NGC4631. Total radio intensity (*contours*) and B-vector at 3.6 cm, observed with the Effelsberg telescope (Krause 2009). The background optical image is from the Misti Mountain Observatory

The exceptionally large radio halos around the irregular and interacting galaxies M82 (Reuter et al. 1992) and NGC4631 (▶ Fig. 13-42) exhibit extremely X-shaped halo fields with almost radial orientations in their inner regions. This indicates that the wind transport is more efficient here than in spiral galaxies. The small gravitational potential of irregular galaxies or external forces by neighboring galaxies may be responsible for high outflow velocities. The mean-field dynamo cannot operate under such conditions. The radio halos of M82 and NGC4631 were resolved into a few magnetic spurs, emerging from star-forming regions in the disk (Golla and Hummel 1994). These observations also support the idea of a fast galactic outflow which is driven by regions of star-formation activity in the disk.

Polarization “vectors” do not distinguish between halo fields which are sheared into elongated loops or regular dynamo-type fields. A large-scale regular field can be measured only by Faraday rotation measures (RM) (▶ Sect. 2.4). RM patterns are very hard to measure in halos because the field components along the line of sight are small. The detailed analysis of the multifrequency observations of the highly inclined galaxy NGC253 (▶ Fig. 13-39) allowed to identify an axisymmetric disk field with even symmetry and an X-shaped halo field, also of *even* symmetry (▶ Fig. 13-43). The polarization asymmetry along the major axis observed at 20 cm in all spiral galaxies with less than 60° inclination observed so far gives further evidence that galaxies host even-parity fields (Braun et al. 2010; Urbanik et al. 1997).

Mean-field dynamo models for galaxies predict disk and halo fields of even symmetry, for which the poloidal (halo) component is of quadrupolar shape (▶ Sect. 2.6). The vertical component of quadrupolar fields is largest near the rotation axis and decreases with distance from the rotation axis. Such an effect is seen in NGC4631 (▶ Fig. 13-42) where the quadrupolar field probably dominates. In several other edge-on galaxies, the vertical field component *increases* with increasing distance from the rotation axis; indicating a superposition of toroidal and poloidal field components. The field strength of pure quadrupoles decreases rapidly with distance R from the center (e.g., Prouza and Šmída 2003), while the observed radial profiles of



■ Fig. 13-43

Model of the symmetric (outwards-directed) halo field of NGC253. The spiral disk field is also symmetric with respect to the plane (from Heesen et al. 2009b) (reproduced with permission © ESO)

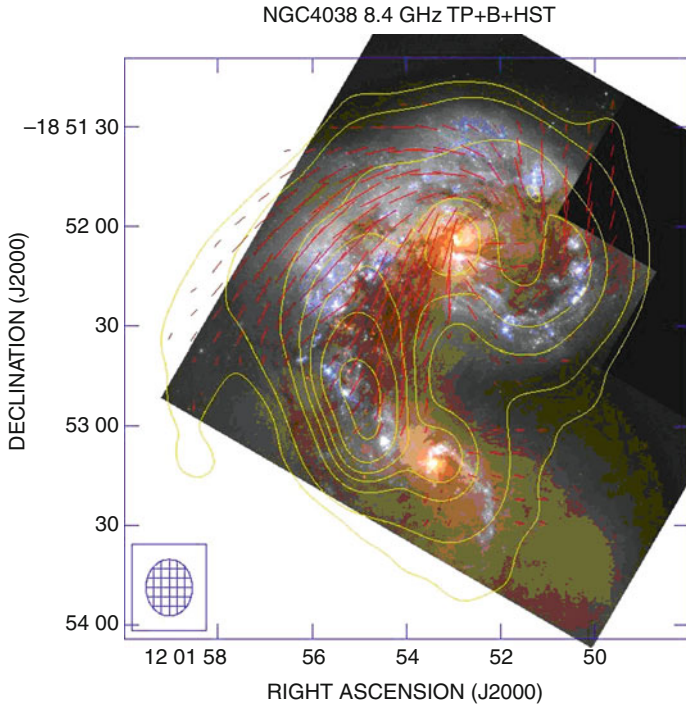
polarized emission show a slow exponential decrease. Dynamo models, including winds, can generate X-shaped fields (▶ Sect. 2.6), which decrease slowly with R .

In summary, the detection of X-shaped fields in all galaxies observed so far can be explained by dynamo action and/or outflows. If outflows are a general phenomenon in galaxies, they can magnetize the intergalactic medium (IGM). Starburst dwarf galaxies in the early Universe were especially efficient in magnetizing the IGM. The extent of magnetic fields into the IGM is not yet visible. Energy losses of the cosmic-ray electrons prevent the emission of radio waves beyond some height, while magnetic fields may still exist much further outward. Low-energy electrons live longer, can propagate further into the IGM and, emit synchrotron emission at low frequencies (▶ Sect. 2.2). Observations with the Low Frequency Array (LOFAR) should reveal much larger radio halos (▶ Sect. 5).

4.8 Interacting Galaxies

Gravitational interaction between galaxies leads to asymmetric gas flows, compression, shear, enhanced turbulence, and outflows. Compression and shear of gas flows can also modify the structure of galactic and intergalactic magnetic fields. In particular, fields can become aligned along the compression front or perpendicular to the velocity gradients. Such gas flows make turbulent fields highly anisotropic.

The classical interacting galaxy pair is NGC4038/39, the “Antennae” (▶ Fig. 13-44). It shows bright, extended radio emission filling the volume of the whole system, with no dominant nuclear sources. In the interaction region between the galaxies, where star formation did not yet start, and at the northeastern edge of the system, the magnetic field is partly ordered, probably the result of compression and shearing motions along the tidal tail, respectively. Particularly strong, almost unpolarized emission comes from a region of violent star formation, hidden in dust, at the southern end of a dense cloud complex extending between the galaxies. In this region, highly turbulent magnetic fields reach strengths of $\approx 30 \mu\text{G}$. The mean total magnetic



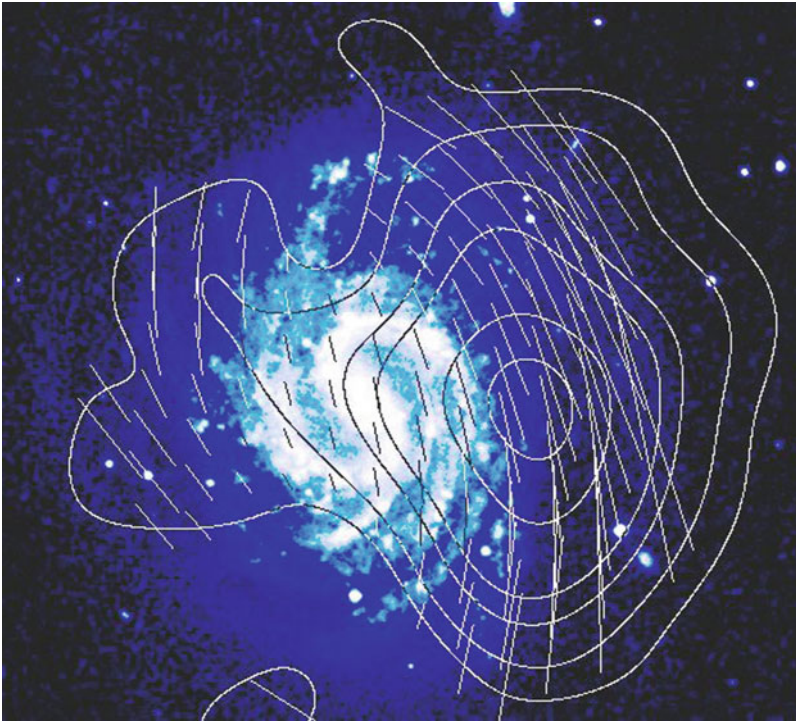
■ Fig. 13-44

“Antennae” galaxy pair NGC4038/39. Polarized radio intensity (*contours*) and B-vectors at 3.6 cm, combined from observations with the VLA and Effelsberg telescopes (Chyży and Beck 2004). The background optical image is from the Hubble Space Telescope

field is stronger than in normal spirals, but the mean degree of polarization is unusually low, implying that the ordered field, generated by compression, has become tangled in the region with violent star formation.

Interaction with a dense intergalactic medium also imprints unique signatures onto magnetic fields and thus the radio emission. The Virgo cluster is a location of especially strong interaction effects, and almost all cluster galaxies observed so far show asymmetries of their polarized emission (☉ Table 13-7 in the Appendix). In NGC4254, NGC4522, and NGC4535 (☛ Fig. 13-45), the polarized emission on one side of the galaxy is shifted toward the edge of the spiral arm, an indication for shear by tidal tails or ram pressure by the intracluster medium. The heavily disrupted galaxy NGC4438 (Vollmer et al. 2007) has almost its whole radio emission (total power and polarized) displaced toward the giant elliptical M86 to which it is also connected by a chain of $H\alpha$ -emitting filaments.

Interaction may also induce violent star-formation activity in the nuclear region or in the disk which may produce huge radio lobes due to outflowing gas and magnetic field. The lobes of the Virgo spiral NGC4569 reach out to at least 24 kpc from the disk and are highly polarized (☛ Fig. 13-46). However, there is neither an active nucleus nor a recent starburst in the disk so that the radio lobes are probably the result of nuclear activity in the past.



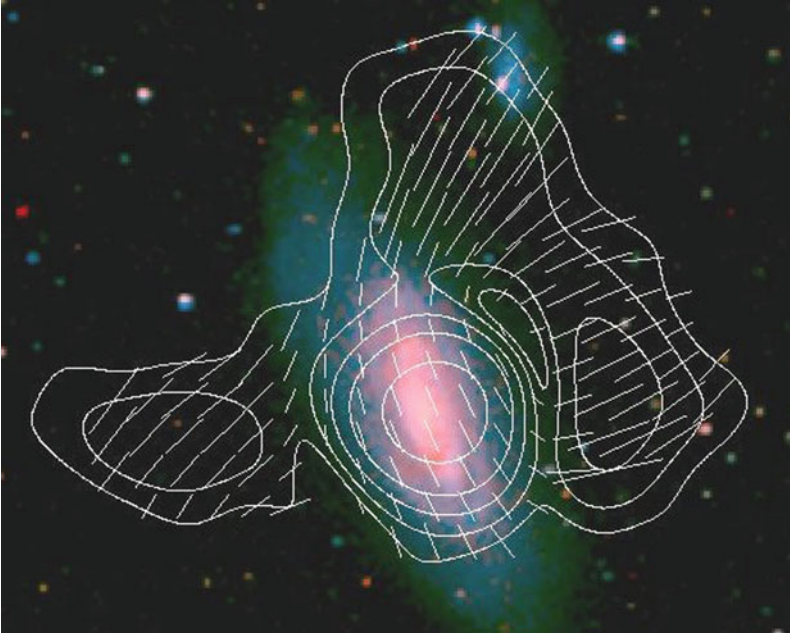
■ Fig. 13-45

Spiral galaxy NGC4535 in the Virgo cluster. Polarized radio intensity (*contours*) and B-vectors at 6 cm, observed with the Effelsberg telescope (Weżgowiec et al. 2007). The background optical image is from the Digital Sky Survey

Tidal interaction is also the probable cause of the asymmetric appearance of NGC3627 within the Leo Triplet (► Fig. 13-47). While the ordered field in the western half is strong and precisely follows the dust lanes, a bright magnetic arm in the eastern half crosses the optical arm and its massive dust lane at a large angle. No counterpart of this feature was detected in any other spectral range. Either the optical arm was recently deformed due to interaction or ram pressure, or the magnetic arm is an out-of-plane feature generated by interaction.

In a few cases, a radio and gaseous bridge has been found between colliding galaxies. The radio emission is due to relativistic electrons pulled out from the disks together with gas and magnetic fields. This phenomenon (called “taffy galaxies”) seems to be rare because only two objects, UGC12914/5 and UGC813/6, were found so far (Condon et al. 2002; Drzazga et al. 2011). This may be due to the steep spectrum of the bridges, making them invisible at centimeter wavelengths in weaker objects.

In compact galaxy groups, tidal interactions may trigger rapid star formation in one or more member galaxies, causing supersonic outflows of hot gas. Some compact groups have long HI tails, indicating strong, tidally-driven outflows of the neutral gas from the system. If the expelled gas was magnetized, it might provide the supply of magnetic fields into the intergalactic space.



■ Fig. 13-46

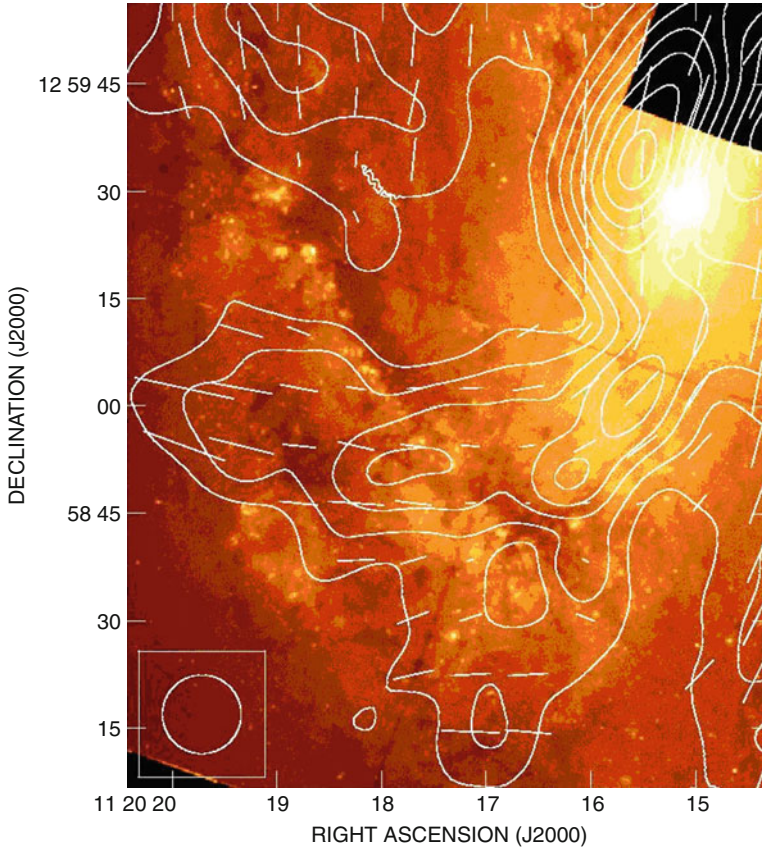
Spiral galaxy NGC4569 in the Virgo cluster. Polarized radio intensity (*contours*) and B-vectors at 6 cm, observed with the Effelsberg telescope (Chyzy et al. 2006). The background optical image is from the Digital Sky Survey

Starburst galaxies (either dwarf and massive) constitute the basic source responsible for the enrichment of the intragroup medium with relativistic particles and magnetic fields. There are grounds to expect that the compact galaxy groups show diffuse radio emission, with a spectrum rapidly steepening away from the cosmic-ray sources in galactic disks.

The best studied example of a compact group is Stephan's Quintet (at a distance of 85 Mpc), with its pool of hot gas extending between the galaxies. It shows a huge, long filament visible in radio continuum. Strong polarization of this intragroup emission (► Fig. 13-48) indicates a substantial content of ordered (probably shock-compressed) magnetic fields.

In summary, polarized radio emission is an excellent tracer of tidal effects between galaxies and of ram pressure in the intracluster medium. As the decompression and diffusion timescales of the field are very long, it keeps memory of events in the past, up to the lifetime of the illuminating cosmic-ray electrons. Low-frequency radio observations will trace interactions which occurred many Gyr ago and are no longer visible in other spectral ranges. Tidal tails from interacting galaxies may also constitute a significant source of magnetic fields in the intracluster and intergalactic media.

NGC3627 3.6cm VLA + Effelsberg Polarized Intensity + B-Vectors HPBW = 11"



■ Fig. 13-47

Interacting spiral galaxy NGC3627. Polarized radio intensity (*contours*) and B-vectors at 3.6 cm, combined from observations with the VLA and Effelsberg telescopes (Soida et al. 2001). The background optical image is from the Hubble Space Telescope

4.9 Galaxies with Jets

Nuclear jets are observed in several spiral galaxies. These jets are weak and small compared to those of radio galaxies and quasars. Detection is further hampered by the fact that they emerge at some angle with respect to the disk so that little interaction with the ISM occurs. Only if the accretion disk is oriented almost perpendicular to the disk, the jet hits a significant amount of ISM matter, cosmic-ray electrons are accelerated in shocks, and the jet becomes radio-bright. This geometry was first proven for NGC4258 by observations of the water maser emission from the accretion disk (Greenhill et al. 1995). This is why NGC4258 is one of the rare cases where a large radio jet of at least 15 kpc length is observed (Krause and Löhner 2004; van Albada and van der Hulst 1982). The total intensity map of NGC4258 (🔗 Fig. 13-49) reveals that the jets

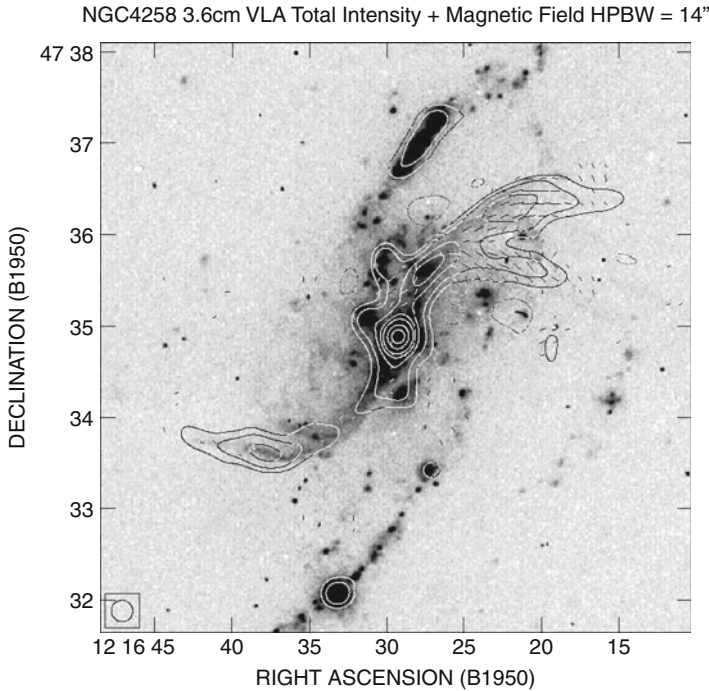


■ Fig. 13-48

Stephan's Quintet of interacting galaxies. Total radio intensity (*contours*) and B-vectors at 6 cm, observed with the VLA (from Marian Soida, Kraków University). The background optical image is from the Hubble Space Telescope

emerge from the Galactic center perpendicular to the accretion disk, which is oriented in east-west direction and is seen almost edge-on, and bend out to become the “anomalous radio arms,” visible out to the boundaries of the spiral galaxy. The magnetic field orientation is mainly along the jet direction. The observed tilt with respect to the jet axis may indicate an additional toroidal field component or a helical field around the jet. The equipartition field strength is about $300 \mu\text{G}$ (at the resolution of about 100 pc), which is a lower limit due to energy losses of the cosmic-ray electrons and the limited resolution.

The barred galaxy NGC7479 also shows remarkable jet-like radio continuum features: bright, narrow, 12 kpc long in projection, and containing an aligned magnetic field (► Fig. 13-50). The lack of any optical or near-infrared emission associated with the jets suggests that at least the outer parts of the jets are extraplanar features, although close to the disk plane. The equipartition strength is $35\text{--}40 \mu\text{G}$ for the total magnetic field and about $10 \mu\text{G}$ for the ordered magnetic field in the jets. According to Faraday rotation measurements, the large-scale regular magnetic field along the bar points toward the nucleus on both sides. Multiple reversals on scales of 1–2 kpc are detected, probably occurring in the galaxy disk in front of the eastern jet by anisotropic fields in the shearing gas flow in the bar potential.



■ Fig. 13-49

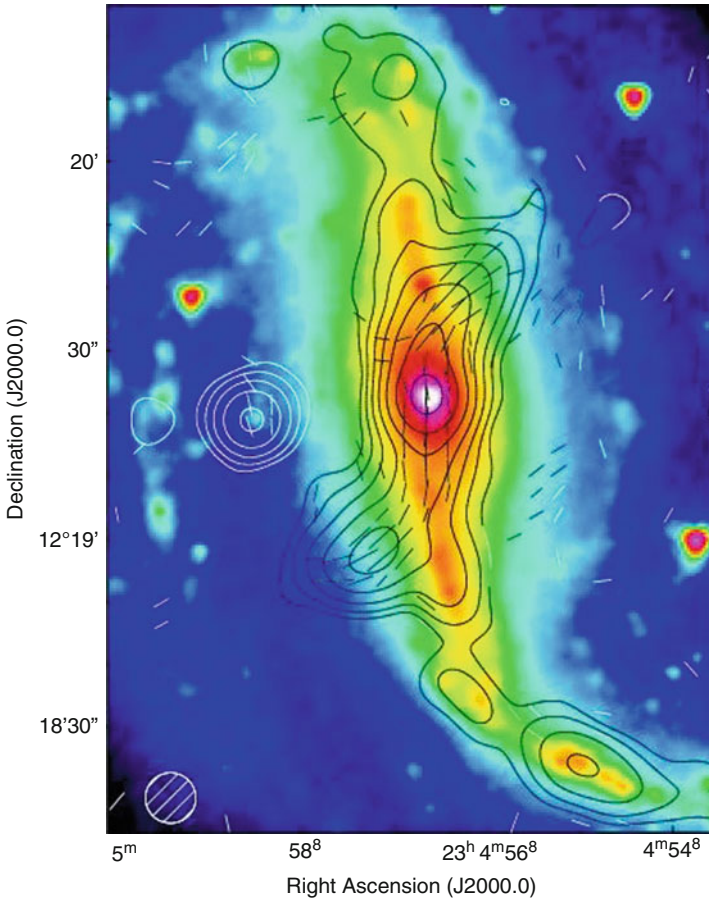
Spiral galaxy NGC4258 with two jets. Total radio intensity (*contours*) and B-vectors at 3.6 cm, observed with the VLA (Krause and Löhr 2004). The background H α image is from the Hoher List Observatory of the University of Bonn

Highly polarized radio emission from kpc-sized jets has also been detected e.g., in NGC3079 (Cecil et al. 2001), with the field orientations perpendicular to the jet's axis), and in the outflow lobes of the Circinus Galaxy (Elmoultie et al. 1995).

Jets in spiral galaxies may be more frequent than the available radio observations suggest. Future low-frequency observations may help because they may show weak synchrotron emission from interface regions between the jets and the low-density halo gas.

4.10 Elliptical and Dwarf Spheroidal Galaxies

Elliptical galaxies with active nuclei are among the brightest known radio sources. Their jets and radio lobes are generated by magneto-hydrodynamic processes which are discussed elsewhere. Radio emission from quiet elliptical and S0 galaxies is also associated with their nuclei (Fabiano et al. 1987). Apart from the nuclear activity, elliptical galaxies are radio-faint because star-formation activity is very low and cosmic-ray electrons are rare. A few ellipticals form stars in their inner regions, but synchrotron emission and hence magnetic fields were not yet detected.



■ Fig. 13-50

Barred spiral NGC7479 with two jets. Total radio intensity (*contours*) and B-vectors at 3.5 cm, observed with the VLA (Laine and Beck 2008). The background shows a Spitzer/IRAC 3.6 μm image (NASA/JPL-Caltech/Seppo Laine)

The existence of magnetic fields in the halos of non-active ellipticals is a matter of speculation. Regular fields are not expected in ellipticals because the lack of ordered rotation prevents the action of the mean-field dynamo. Dwarf spheroidal galaxies have some ordered rotation, but lack turbulent gas. Turbulence in the hot gas of large ellipticals may drive a small-scale dynamo and generate turbulent fields with a few μG strength and turbulent scales of a few 100 pc (Moss and Shukurov 1996). However, there are no cosmic-ray electrons and, hence, no synchrotron emission. Detection of turbulent magnetic fields is only possible via the dispersion of Faraday rotation measures toward polarized background sources. Most large ellipticals are located in galaxy clusters where Faraday rotation will be dominated by the turbulent fields of the intracluster gas. For small ellipticals, the number of polarized background sources will only

be sufficient with much more sensitive radio telescopes like the SKA. This leaves only isolated giant ellipticals for future studies.

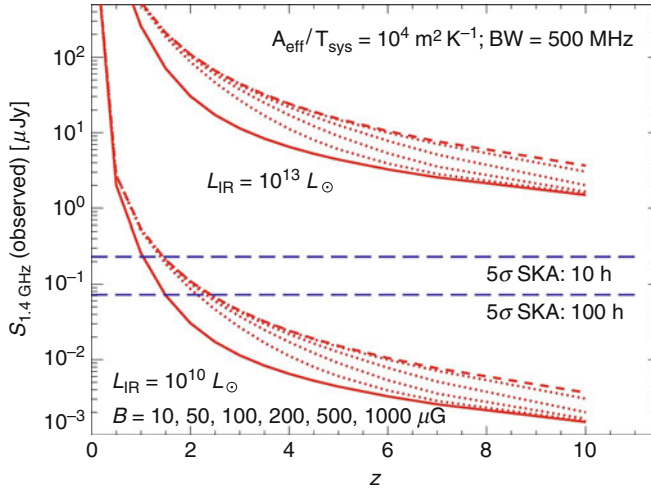
Dwarf spheroidal galaxies are of interest to search for synchrotron emission from secondary electrons and positrons generated by the decay of dark-matter by WIMP annihilations, e.g., neutralinos (Colafrancesco et al. 2007). These galaxies do not generate thermal emission or primary electrons from star formation. Detection of radio emission would be of high importance, but all attempts failed so far. The main uncertainty is origin of magnetic fields in such systems (see above). Only if the field strength is a few μG , detection of synchrotron emission from dark-matter decay may be possible.

5 Outlook

Thanks to radio polarization observations, the global properties of interstellar magnetic fields in external galaxies and the field structures on pc and sub-pc sizes in the Milky Way are reasonably well known. However, the processes connecting the features at large and small scales are not understood because the *angular* resolution in external galaxies is too low with present-day radio telescopes. Most of the existing polarization data are observed in wide frequency bands and hence suffer from very low *spectral* resolution, which causes depolarization by gradients of Faraday rotation or by different Faraday rotation components within the beam or along the line of sight. Modern radio telescopes are (or will be) equipped with multi-channel polarimeters, allowing application of *RM Synthesis* (➤ Sect. 2.4) and resolving RM components along the line of sight. This method is going to revolutionize radio polarization observations.

New and planned telescopes will widen the range of observable magnetic phenomena. The importance of polarimetry for the planned giant optical telescopes still needs to be established, while huge progress is expected in the radio range. The PLANCK satellite and several balloon instruments (PILOT, BLAST-pol) will improve the sensitivity of polarimetry in the sub-millimeter range at arcminute resolution. The Atacama Large Millimeter Array (ALMA) will provide greatly improved sensitivity at arcsecond resolution for detailed imaging diffuse polarized emission from dust grains and for detection of the Zeeman effect in molecular clouds. High-resolution, deep observations at high frequencies (≥ 5 GHz), where Faraday effects are small, require a major increase in sensitivity for continuum observations which will be achieved by the Extended Very Large Array (EVLA) and the planned Square Kilometre Array (SKA). The detailed structure of the magnetic fields in the ISM of galaxies and in galaxy halos will be observed, giving direct insight into the interaction between magnetic fields and the various gas components. High angular resolution is also needed to distinguish between regular and anisotropic (sheared) fields and to test various models of the interaction between spiral shocks and magnetic fields. The power spectra of turbulent magnetic fields could be measured down to small scales. The SKA will also allow to measure the Zeeman effect in much weaker magnetic fields in the Milky Way and in nearby galaxies.

The SKA will detect synchrotron emission from Milky Way-type galaxies at redshifts of $z \leq 1.5$ (➤ Fig. 13-51) and their polarized emission to $z \leq 0.5$ (assuming 10% polarization). Bright starburst galaxies could be observed at larger redshifts, but are not expected to host ordered or regular fields. Total synchrotron emission, signature of total magnetic fields, could be detected with the SKA out to large redshifts for starburst galaxies, depending on luminosity



■ Fig. 13-51

Total synchrotron emission at 1.4 GHz as a function of redshift z , total magnetic field strength B and total infrared luminosity L_{IR} . The 5σ detection limits for 10 and 100 h integration time with the SKA are also shown (Murphy 2009)

and magnetic field strength (● Fig. 13-51). However, for fields weaker than $3.25 \mu\text{G} (1+z)^2$, energy loss of cosmic-ray electrons is dominated by the inverse Compton effect with photons of the cosmic microwave background so that their energy is transferred mostly to the X-ray and not to the radio domain. On the other hand, for strong fields, the energy range of electrons emitting in the GHz range shifts to low energies, where ionization and bremsstrahlung losses become dominant. The mere detection of synchrotron emission from galaxies at high redshifts will constrain the range of allowed magnetic field strengths.

Dynamo theory predicts timescales of amplification and coherent ordering of magnetic fields in galaxies (● Sect. 2.6). Based on models describing the formation and evolution of dwarf and disk galaxies, the probable evolution of turbulent and regular magnetic fields can be tested observationally (Arshakian et al. 2009):

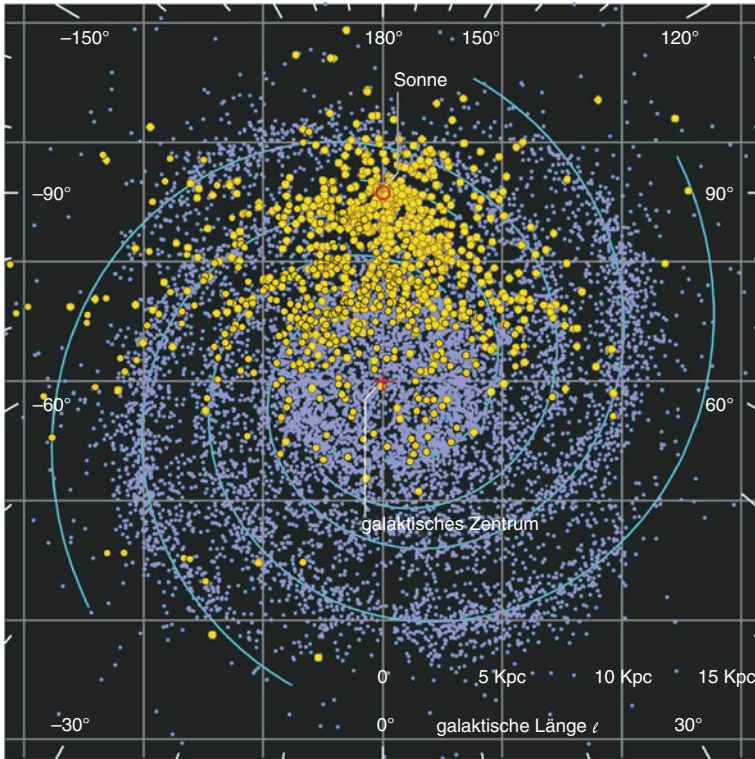
- Strong turbulent fields (in equipartition with turbulent gas motions) and, hence, unpolarized synchrotron emission are expected in galaxies at $z < 10$.
- Strong regular fields (which are coherent over a scale of about 1 kpc) and, hence, polarized synchrotron emission and fluctuating RMs are expected in galaxies at $z \leq 3$.
- Large-scale patterns of fully coherent regular fields and, hence, polarized synchrotron emission and large-scale RM patterns are expected in dwarf and MilkyWay-type galaxies at $z \leq 1$.
- Giant galaxies (disk radius > 15 kpc) have not yet generated fully coherent fields.
- Major mergers enhanced turbulent fields, but destroyed regular fields and delayed the formation of fully coherent fields. The lack of regular fields in nearby galaxies can be a signature of major mergers in the past.

The detections of total synchrotron emission in starburst galaxies at $z \leq 4$ and of RMs from intervening galaxies at $z \leq 2$ (Sect. 4.2) are consistent with dynamo theory. Observed field patterns are so far in agreement with the predictions of the mean-field dynamo (► Sects. 4.2 and ► 4.7). Progress is needed in numerical MHD simulations. Crucial tests of dynamo action will be possible in young galaxies. Detection of regular fields at $z \geq 3$ would call for a faster process than the dynamo. On the other hand, the failure to detect global coherent field patterns in galaxies $z \leq 1$ would indicate that the time needed for field ordering is even longer than the mean-field dynamo theory predicts, or that this theory is not applicable. If bisymmetric spiral (BSS) magnetic patterns turn out to dominate, in contrast to nearby galaxies, this would indicate that the fields could be primordial or intergalactic fields which are twisted and amplified by differential rotation.

If polarized emission of galaxies is too weak to be detected, the method of *RM grids* toward background QSOs could still be applied to measure the strength and structure of regular fields. The accuracy is determined by the polarized flux of the background QSO which could be much higher than that of the intervening galaxy. A reliable model for the structure of the magnetic field of nearby galaxies needs many RM values, hence a sufficiently large number density of polarized background sources, calling for high sensitivity. Faraday rotation in the direction of QSOs could even be measured in galaxies at distances near to those of young QSOs ($z \approx 5$). The RM values are reduced by the redshift dilution factor of $(1+z)^{-2}$ so that high RM accuracy is needed.

The SKA will be able to detect $1 \mu\text{Jy}$ sources and measure about 10^4 RMs per square degree at 1.4 GHz within 12-h integration time. The SKA Magnetism Key Science Project plans to observe an all-sky RM grid with 1-h integration per field (Gaensler et al. 2004) which should contain about 20,000 RMs from pulsars in the Milky Way with a mean spacing of $30'$ (► Fig. 13-52) and several 100 extragalactic pulsars. At least 10^7 RMs from compact polarized extragalactic sources at a mean spacing of about $1.5'$ are expected, about 10,000 in the area around M31 (► Fig. 13-53). This fundamental survey will be used to model the structure and strength of the magnetic fields in the foreground, i.e., in the Milky Way, in intervening galaxies, and in the intergalactic medium. A pilot all-sky survey called POSSUM is planned with the Australian SKA Precursor (ASKAP). MeerKAT, the South African SKA precursor, and APERTIF, the Dutch SKA pathfinder telescope, will have a higher sensitivity but a smaller field of view and will concentrate on measuring RM grids centered on individual objects.

Progress is also expected at low radio frequencies. Present-day measurements of galactic magnetic fields by synchrotron emission are limited by the lifetime and diffusion length of the cosmic-ray electrons which illuminate the fields. With typical diffusion lengths of only 1 kpc away from the acceleration sites in star-forming regions, the size of galaxies at centimeter wavelengths is not much larger than that in the optical or infrared spectral ranges. There is indication that magnetic fields probably extend much further into the intergalactic space (► Sect. 4.7). The Low Frequency Array (LOFAR), and the planned radio telescopes Murchison Widefield Array (MWA) and Long Wavelength Array (LWA), and the low-frequency part of the SKA will be suitable instruments to search for extended synchrotron radiation at the lowest possible levels in outer galaxy disks and halos and investigate the transition to intergalactic space. While most of the disk is depolarized at low frequencies, polarization should be detectable from the outer regions. Faraday rotation in the Earth's ionosphere and in the Milky Way foreground is strong and need to be corrected for.



■ Fig. 13-52

Simulation of pulsars in the Milky Way that will be detected with the SKA (blue), compared to about 2,000 pulsars known today (yellow) (from Jim Cordes, Cornell University) (Graphics: Sterne und Weltraum)

The filaments of the local Cosmic Web may contain *intergalactic magnetic fields*, possibly enhanced by IGM shocks, and this field may be detectable by direct observation of total synchrotron emission or by Faraday rotation toward background sources. For fields of $10^{-8} - 10^{-7}$ G with 1 Mpc coherence length and 10^{-5} cm^{-3} electron density, $|\text{RM}|$ of $0.1 - 1$ rad m^{-2} are expected. An overall intergalactic field is much weaker and may only become evident as increased RMs toward QSOs at redshifts of $z > 3$ by averaging over a large number of sources. As the Faraday rotation angle increases with λ^2 , searches for low $|\text{RM}|$ should preferably be done at low frequencies.

In summary, the SKA and its pathfinders (APERTIF, EVLA, LOFAR, LWA, MWA) and precursors (ASKAP and MeerKAT) will measure the structure and strength of the magnetic fields in the Milky Way, in intervening galaxies, and possibly in the intergalactic medium. Looking back into time, the future telescopes could shed light on the origin and evolution of cosmic magnetic fields. The observational methods are:

- RM grids of extragalactic sources and pulsars to map the detailed 3D structure of the Milky Way's magnetic field (0.2–1 GHz)



■ Fig. 13-53

Simulation of RMs toward background sources (*white points*) in the region of M31 observable with the SKA within 1 h integration time. Optical emission from M31 is shown in *red*, diffuse total radio continuum intensity in *blue*, and diffuse polarized intensity in *green* (from Bryan Gaensler, Sydney University)

- High-resolution mapping of total and polarized synchrotron emission from galaxy disks and halos of nearby galaxies at high frequencies (≥ 5 GHz)
- Mapping of the total and polarized synchrotron emission from the outer disks and halos of nearby galaxies and galaxy groups at low frequencies (≤ 0.3 GHz)
- Reconstruction of 3D field patterns in nearby galaxies by RM Synthesis of the diffuse polarized emission
- Reconstruction of 3D field patterns in nearby galaxies from RMs toward polarized background sources
- Recognition of simple patterns of regular fields in galaxies from RMs toward polarized background sources (at $z \leq 0.02$)
- Search for polarized synchrotron emission from distant galaxies (at $z \leq 0.5$)
- Search for total synchrotron emission from distant starburst galaxies (at $z \leq 3$)
- Search for regular fields in very distant intervening galaxies toward QSOs (at $z \leq 5$)

Fundamental questions are waiting to be answered:

- When were the first magnetic fields generated: in young galaxies, in protogalactic clouds, or are they relics from the early Universe before the galaxies were formed?
- How and how fast were magnetic fields amplified in the interstellar medium?
- Did magnetic fields affect the evolution of galaxies?
- How important are magnetic fields for the physics of galaxies, like the efficiency to form stars from gas, the formation of spiral arms or the generation of outflows?
- Can outflows from galaxies magnetize the intergalactic space?
- How strong and how ordered are magnetic fields in intergalactic space?
- What is the large-scale structure of the Milky Way's magnetic field?
- How strongly are extragalactic ultrahigh-energy cosmic rays deflected in the Milky Way and in intergalactic space?

Acknowledgments

The authors would like to thank many of our colleagues who have pursued the studies of magnetic fields in the Milky Way and in galaxies for the past 40 years, especially Wolfgang Reich, Marita Krause and Patricia Reich at MPIfR. Many excellent cooperation projects in this field were performed with groups in Kraków (Poland), DRAO Penticton (Canada), NAOC Beijing (China), Moscow and Perm (Russia), Newcastle (UK), Potsdam and Bochum (Germany). Marek Urbanik is acknowledged for compiling [Table 13-7](#). Elly M. Berkhuijsen and Anvar Shukurov are acknowledged for careful reading of the manuscript.

Appendix

A.1 Catalogue of Radio Polarization Observations of Nearby Galaxies

In radio continuum the typical degrees of polarization are much higher than those in the other spectral ranges, and further benefit comes from the development of large instruments and sensitive receivers. This is why almost all of our knowledge on interstellar magnetic fields in galaxies is based on their polarized radio emission and Faraday rotation.

A list of spiral, barred, irregular, and dwarf galaxies detected in radio polarization until year 2010 is given in [Tables 13-5–13-7](#). Most detections were made in the wavelength range 2–13 cm, where Faraday depolarization is small. At ≈ 20 cm, the polarized intensity is generally smaller by a factor of several ([Fig. 13-3](#)). At wavelengths of more than 20 cm, no detection of polarized emission from spiral galaxies has been reported so far.

Table 13-5

Radio polarization observations and magnetic field structures of galaxies with low or moderate inclination

Galaxy	Telescope & λ (cm)	Structure	Reference
M33	E21,18,11,6,3	BSS	Buczilowski and Beck 1991
	E6,4, V21	ASS+BSS	Tabatabaei et al. 2008
M51	W21,6	Spiral	Segalovitz et al. 1976
	V21,18	Spiral	Horellou et al. 1992
	E6,3, V21,6	ASS+BSS (disk)+ASS (halo)	Berkhuijsen et al. 1997 ; Neininger 1992
	W21	BSS (halo)	Heald et al. 2009
	E6,4, V21,6,4	ASS+QSS (disk)+BSS (halo)	Fletcher et al. 2011
M81	E6, V21	BSS?	Krause et al. 1989b
	E6, V21	BSS (+ASS)?	Sokoloff et al. 1992
	V21,6	Interarm fields	Schoofs 1992

■ Table 13-5
(Continued)

Galaxy	Telescope & λ (cm)	Structure	Reference
M83	V21	Spiral	Sukumar and Allen 1989
	P6	Spiral	Harnett et al. 1990
	E6,3	BSS?	Neininger et al. 1991, 1993
	A13	Magnetic arms	Ehle 1995
	V6	Magnetic arms + bar	Beck 2005
M101	E11,6,3	Spiral	Gräve et al. 1990
	E11,6	Spiral	Berkhuijsen et al. in prep.
NGC0628	W21	Incomplete spiral	Heald et al. 2009
NGC0660	V6	polar ring + X-shape	Drzazga et al. 2011
NGC0877	V6	Spiral	Drzazga et al. 2011
NGC1097	V21,18,6,4	ASS+BSS+QSS+ bar+nuclear spiral	Beck et al. 2005a
NGC1365	V21,18,6,4	ASS+BSS+QSS+ bar+nuclear spiral	Beck et al. 2005a
NGC1559	A13,6	bar	Beck et al. 2002
NGC1566	A21,13,6	Spiral, interarm	Ehle et al. 1996
NGC1569	W21, V6,3	Spiral, bubbles, loops	Kepley et al. 2010
NGC1672	A13,6	Spiral, interarm	Beck et al. 2002
NGC2207	V6	Spiral + radial streamers	Drzazga et al. 2011
NGC2276	V21,6	BSS?	Hummel and Beck 1995
NGC2403	E11,6	Spiral	Beck unpubl.
	W21	Diffuse	Heald et al. 2009
NGC2442	A13,6	Spiral + bar	Harnett et al. 2004
NGC2841	W21	Two arcs	Heald et al. 2009
NGC2903	E6,3, V21	Spiral	Beck unpubl.
	W21	Spiral	Heald et al. 2009
NGC2997	V21,6,4, A13	Spiral +inner ASS?	Han et al. 1999
NGC3521	E3	Spiral, compressed	Knapik et al. 2000
NGC3627	E3	Spiral + dust lane	Soida et al. 1999
	V6,4	Anomalous arm	Soida et al. 2001
	W21	Spiral	Heald et al. 2009
NGC3938	W21	Spiral	Heald et al. 2009
NGC4038	V21,6,4	tidal arm	Chyży and Beck 2004
NGC4214	V6	No ordered field	Kepley et al. 2009
	E6	Fragment of a spiral	Drzazga 2008
NGC4258	W21, V21	In anomalous arms	van Albada and van der Hulst 1982
	V21,6	anomalous arms	Hummel et al. 1989
	V4, E3	nuclear jet + an. arms	Krause and Löhner 2004
NGC4414	V6,4	ASS+BSS+QSS	Soida et al. 2002
NGC4449	E6,3	opt. filaments	Klein et al. 1996
	V6,4	Spiral+radial field	Chyży et al. 2000
NGC4490/85	E6,4	Radial halo field	Knapik et al. in prep.
NGC4736	V6,4	Spiral, ASS?	Chyży and Buta 2008
	W21	Outer lobe	Heald et al. 2009

■ **Table 13-5**
(Continued)

Galaxy	Telescope & λ (cm)	Structure	Reference
NGC5033	W21	Inner disk	Heald et al. 2009
NGC5055	E3	Spiral	Knapik et al. 2000
	W21	Spiral	Heald et al. 2009
NGC5426/7	V6	Spiral + spiral	Drzazga et al. 2011
NGC6822	E11,6,3	No ordered field	Chyży et al. 2003, 2011
NGC6907	V6	Spiral	Drzazga et al. 2011
NGC6946	E11,6,3	ASS?	Ehle and Beck 1993
	V21,18,6,4	ASS+QSS	Beck 1991, 2007
	W21	ASS (halo)	Heald et al. 2009
NGC7479	V21,6,4	spiral jet	Beck et al. 2002, Laine and Beck 2008
NGC7552	A6	Spiral + bar	Beck et al. 2002
IC10	E11,6,3	H α filament	Chyży et al. 2003, 2011
	V6	Filaments	Heesen et al. 2011a
	V6,4	Filaments	Chyży 2005
IC342	E11,6	ASS	Gräve and Beck 1988
	E6, V21	ASS	Krause et al. 1989a
	E6, V21	ASS	Sokoloff et al. 1992
	V6,4	Magnetic arms	Beck unpubl.
IC1613	E11,6	No ordered field	Chyży et al. 2011
UGC813/6	V6	\perp bridge	Drzazga et al. 2011
UGC12914/5	V6	bridge	Drzazga et al. 2011
Holmberg II	E11,6	No ordered field	Chyży et al. 2011
SMC	P21,13	main ridge	Haynes et al. 1986
	A21	Pan-Magellanic?	Mao et al. 2008
LMC	P21,13,6	Magn. loop near 30 Dor	Haynes et al. 1991; Klein et al. 1993
	A21	ASS	Gaensler et al. 2005
Stephan's Quintet	E6,4, V6	Intergalactic field	Soida et al. unpubl.
PKS1229-021	V21,6,2	BSS?	Kronberg et al. 1992

■ **Table 13-6**

Radio polarization observations and magnetic field structures of galaxies with high inclination (almost edge-on)

Galaxy	Telescope & λ (cm)	Structure	Reference
M31	E21,11,6	Even ASS	Beck 1982; Beck et al. 1989
	V21,6	Spiral (inner region)	Beck et al. 1998
	E11,6, V21	Even ASS (+QSS)	Berkhuijsen et al. 2003; Fletcher et al. 2004
M82	V6,4	Radial halo field	Reuter et al. 1994
	E1	disk + vertical halo field	Wielebinski 2006
M104	V21,6	disk+X-shaped halo field	Krause et al. 2006

■ Table 13-6
(continued)

Galaxy	Telescope & λ (cm)	Structure	Reference
NGC253	P6,3	plane	Harnett et al. 1990
	V21,6	plane	Carilli et al. 1992
	E6,3	plane, halo spurs	Beck et al. 1994
	E6,4, V21,6	Even ASS disk field+even halo field	Heesen et al. 2009a, b
	V21,6,4	Helical field in outflow cone	Heesen et al. 2011b
NGC891	V6	plane +halo spurs	Sukumar and Allen 1991
	V21	plane +halo spurs	Hummel et al. 1991
	E3	plane +tilted	Dumke et al. 1995
	E4	Even ASS disk field+X-shaped halo field	Krause 2009
NGC1808	V21,6	Halo spurs	Dahlem et al. 1990
NGC3079	V6	Extraplanar jet	Duric and Seaquist 1988
NGC3432	E6	Vertical, weak	Drzazga 2008
NGC3628	V21	Fragments of ord. field	Reuter et al. 1991
	E3	plane	Dumke et al. 1995
	E4	plane+X-shaped halo field	Krause unpubl.
NGC4217	V6	X-shaped halo field	Soida 2005
NGC4236	E6	No ordered field	Chyży et al. 2007
NGC4565	V21	plane	Sukumar and Allen 1991
	E3	plane	Dumke et al. 1995
	E6,4, V6	plane+X-shaped halo field	Krause 2009
NGC4631	V21	⊥plane	Hummel et al. 1991
	V6,4	⊥plane, spurs	Golla and Hummel 1994
	V21,18	X-shaped halo field	Beck 2009
	E4	plane +vertical central field+X-shaped halo field	Krause 2009
	W21	X-shaped halo field	Heald et al. 2009
NGC4656	E6	No ordered field	Chyży et al. 2007
NGC4666	V21,6	X-shaped halo field	Dahlem et al. 1997
	V6	X-shaped halo field	Soida 2005
NGC4945	P6,3	Halo spurs	Harnett et al. 1989, 1990
NGC5775	V21,6	X-shaped halo field	Tüllmann et al. 2000
	V4	Even ASS disk field+X-shaped halo field	Soida et al. unpubl.
NGC5907	E6, V21	plane	Dumke 1997
	E4	plane +X-shape?	Krause unpubl.
NGC7331	E3	plane	Dumke et al. 1995
	W21	X-shaped halo field	Heald et al. 2009
Circinus	A21, 13	radio lobes	Elmouttie et al. 1995
IC2574	E6	No ordered field	Chyży et al. 2007

■ Table 13-7

Radio polarization observations and magnetic field structures of galaxies in the Virgo cluster

Galaxy	Telescope & λ (cm)	Structure	Reference
NGC4192	E6,4	ASS? +halo field	Weżgowiec et al. 2012
	V21,6	disk +inclined	Vollmer et al. in prep.
NGC4254	E6,3	Spiral	Soida et al. 1996
	V21,6,4, E6,3	ASS (+BSS), tidally stretched	Chyży 2008
	W21	Spiral	Heald et al. 2009
NGC4294	V21,6	Halo field, inclined to disk	Vollmer et al. in prep.
NGC4298	V21,6	Mostly disk	Vollmer et al. in prep.
NGC4299	V21,6	Fragments of a spiral	Vollmer et al. in prep.
NGC4302/NGC4298	E6,4	disk+intergal. bridge, locally vertical	Weżgowiec et al. 2012
NGC4303	E6,4	ASS?	Weżgowiec et al. 2012
	V21,6	Spiral	Vollmer et al. in prep.
NGC4321	V21,6	Spiral	Vollmer et al. 2007, 2010
	W21	Spiral	Heald et al. 2009
	E6,4	BSS? + bar	Weżgowiec et al. 2012
NGC4330	V21,6	Mostly disk	Vollmer et al. in prep.
NGC4388	E6,4	Inclined to disk	Weżgowiec et al. 2012
	V21,6	disk +incl. in halo	Vollmer et al. 2007, 2010
NGC4396	V21,6	disk + NW tail	Vollmer et al. 2007, 2010
NGC4402	V21,6	disk in southern halo, incl. in northern halo	Vollmer et al. 2007, 2010
NGC4419	V21,6	disk +X-shaped	Vollmer et al. in prep.
NGC4438	E6	disk, \perp outflow	Weżgowiec et al. 2007
	V6	disk, displaced from disk in the east	Vollmer et al. 2007, 2010
NGC4457	V21,6	Mostly spiral	Vollmer et al. in prep.
NGC4501	E6,3	disk, asymmetric	Weżgowiec et al. 2007
	V21,6	Compressed along SW disk edge	Vollmer et al. 2007, 2010
NGC4522	V21,6	plane, compressed	Vollmer et al. 2004
NGC4532	V21,6	Huge halo field, inclined+vertical, X-shaped	Vollmer et al. in prep.
NGC4535	V21,6,4	Spiral	Beck et al. 2002
	E6,4	Spiral, asymmetric	Weżgowiec et al. 2007, 2012
	V21,6	spiral arm, asymmetric	Vollmer et al. 2007, in prep.
	E4	ASS?	Weżgowiec et al. 2012
NGC4548	E6	\perp bar	Weżgowiec et al. 2007
NGC4567/NGC4568	V21,6	Intergal. bridge	Vollmer et al. in prep.
NGC4569	E6,4	disk + outflow	Chyży et al. 2006
	W21	disk + outflow	Heald et al. 2009
	V21,6	disk + outflow	Chyży et al. in prep.

■ **Table 13-7**
(Continued)

Galaxy	Telescope & λ (cm)	Structure	Reference
NGC4579	V21,6	Radial inside, spiral in outer disk	Vollmer et al. in prep.
NGC4654	E6,4	SW arm + gas tail	Weżgowiec et al. 2007
	V21,6	arms, bending out toward gas tail	Vollmer et al. 2007, 2010
NGC4689	V21,6	Fragments of a spiral	Vollmer et al. in prep.
NGC4713	V21,6	Spiral	Vollmer et al. in prep.
NGC4808	V21,6	Vertical, asymmetric	Vollmer et al. in prep.

Instruments: *E* Effelsberg 100-m, *V* very large array, *A* Australia telescope compact array, *P* Parkes 64-m, *W* Westerbork synthesis radio telescope

Wavelength code: 21 = 20–22 cm, 18 = 18.0 cm, 13 = 12.5–13.4 cm, 11 = 11.1 cm, 6 = 5.8–6.3 cm, 4 = 3.6 cm, 3 = 2.8 cm, 2 = 2.0 cm, 1 = 9 mm

Field structures: *ASS* axisymmetric spiral, *BSS* bisymmetric spiral, *QSS* quadrisymmetric spiral, *MSS* multimode spiral

A.2 Links to the SKA Project and Its Precursor and Pathfinder Telescopes

<http://www.skatelescope.org>

http://www.scholarpedia.org/article/Square_kilometre_array

<http://www.atnf.csiro.au/SKA>

<http://www.ska.ac.za>

<http://www.aoc.nrao.edu/evla>

<http://www.lofar.org>

<http://www.astron.nl/general/apertif/apertif>

<http://www.phys.unm.edu/~lwa>

<http://www.mwatelescope.org>

Cross-References

- [Astrophysics of Galactic Charged Cosmic Rays](#)
- [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- [Dynamics of Disks and Warps](#)
- [Galactic Distance Scales](#)
- [Gamma-Ray Emission of Supernova Remnants and the Origin of Galactic Cosmic Rays](#)
- [Mass Distribution and Rotation Curve in the Galaxy](#)

References

- | | |
|---|--|
| Appenzeller, I. 1967, <i>PASP</i> , 79, 600 | Athanassoula, E. 1992, <i>MNRAS</i> , 259, 345 |
| Arshakian, T. G., & Beck, R. 2011, <i>MNRAS</i> , 418, 2336 | Axon, D. J., & Ellis, R. S. 1976, <i>MNRAS</i> , 177, 499 |
| Arshakian, T. G., Beck, R., Krause, M., & Sokoloff, D. 2009, <i>A&A</i> , 494, 21 | Balbus, S. A., & Hawley, J. F. 1998, <i>Rev Mod Phys</i> , 70, 1 |

- Battaner, E., & Florido, E. 2000, *Fundam Cosmic Phys*, 21, 1
- Beck, R. 1982, *A&A*, 106, 121
- Beck, R. 1991, *A&A*, 251, 15
- Beck, R. 2001, *Space Sci Rev*, 99, 243
- Beck, R. 2005, in *Cosmic Magnetic Fields*, eds. R. Wielebinski, & R. Beck (Berlin: Springer), 41
- Beck, R. 2007, *A&A*, 470, 539
- Beck, R. 2009, *Astrophys Space Sci*, 320, 77
- Beck, R., & Krause, M. 2005, *AN*, 326, 414
- Beck, R., Loiseau, N., Hummel, E. et al. 1989, *A&A*, 222, 58
- Beck, R., Carilli, C. L., Holdaway, M. A., & Klein, U. 1994, *A&A*, 292, 409
- Beck, R., Brandenburg, A., Moss, D., Shukurov, A., & Sokoloff, D. 1996, *ARAA*, 34, 155
- Beck, R., Berkhuijsen, E. M., & Hoernes, P. 1998, *A&AS*, 129, 329
- Beck, R., Shoutenkov, V., Ehle, M. et al. 2002, *A&A*, 391, 83
- Beck, R., Shukurov, A., Sokoloff, D., & Wielebinski, R. 2003, *A&A*, 411, 99
- Beck, R., Fletcher, A., Shukurov, A. et al. 2005a, *A&A*, 444, 739
- Beck, R., Ehle, M., Fletcher, A. et al. 2005b, in *The Evolution of Starbursts*, AIP Conf. Proc. Vol. 783, ed. S. Hüttemeister et al. (Melville, NY: American Institute of Physics), 216
- Behr, A. 1961, *ZfA*, 53, 95
- Bennett, C. L., Hill, R. S., Hinshaw, G. et al. 2003, *ApJS*, 148, 97
- Berkhuijsen, E. M., & Brouw, W. N. 1963, *BAN*, 17, 185
- Berkhuijsen, E. M., Horellou, C., Krause, M. et al. 1997, *A&A*, 318, 700
- Berkhuijsen, E. M., Beck, R., & Hoernes, P. 2003, *A&A*, 398, 937
- Beuermann, K., Kanbach, G., & Berkhuijsen, E. M. 1985, *A&A*, 153, 17
- Boulares, A., & Cox, D. P. 1990, *ApJ*, 365, 544
- Brandenburg, A., & Subramanian, K. 2005, *Phys Rep*, 417, 1
- Braun, R., Heald, G., & Beck, R. 2010, *A&A*, 514, A42
- Brentjens, M. A., & de Bruyn, A. G. 2005, *A&A*, 441, 1217
- Brouw, W. N., & Spoelstra, T. A. T. 1976, *A&AS*, 26, 129
- Brown, J. C., Haverkorn, M., Gaensler, B. M. et al. 2007, *ApJ*, 663, 258
- Buczilowski, U. R., & Beck, R. 1991, *A&A*, 241, 47
- Burn, B. J. 1966, *MNRAS*, 133, 67
- Carilli, C. L., Holdaway, M. A., Ho, P. T. P., & de Pree, C. G. 1992, *ApJ*, 399, L59
- Carretti, E., Haverkorn, M., McConnell, D. et al. 2010, *MNRAS*, 405, 1670
- Cecil, G., Bland-Hawthorn, J., Veilleux, S., & Filippenko, A. V. 2001, *ApJ*, 555, 338
- Chandrasekhar, S., & Fermi, E. 1953, *ApJ*, 118, 113
- Cho, J., & Lazarian, A. 2005, *ApJ*, 631, 361
- Chyży, K. T. 2005, in *From 30 Doradus to Lyman Break Galaxies*, ed. R. de Grijs and R. M. Gonzales Delgado (Dordrecht: Springer), P12
- Chyży, K. T. 2008, *A&A*, 482, 755
- Chyży, K. T., & Beck, R. 2004, *A&A*, 417, 541
- Chyży, K. T., & Buta, R. J. 2008, *ApJ*, 677, L17
- Chyży, K. T., Beck, R., Kohle, S., Klein, U., & Urbanik, M. 2000, *A&A*, 356, 757
- Chyży, K. T., Knapik, J., Bomans, D. J. et al. 2003, *A&A*, 405, 513
- Chyży, K. T., Soida, M., Bomans, D. J. et al. 2006, *A&A*, 447, 465
- Chyży, K. T., Bomans, D. J., Krause, M. et al. 2007, *A&A*, 462, 933
- Chyży, K. T., Weźgowiec, M., Beck, R., & Bomans, D. J. 2011, *A&A*, 529, A94
- Colafrancesco, S., Profumo, S., & Ullio, P. 2007, *Phys Rev D*, 75, 023513
- Condon, J. J., Helou, G., & Jarrett, T. H. 2002, *AJ*, 123, 1881
- Crocker, R. M., Jones, D. I., Melia, F. et al. 2010, *Nature*, 463, 65
- Crutcher, R. M., Kazes, I., & Troland, T. H. 1987, *A&A*, 181, 119
- Crutcher, R. M., Troland, T. H., Lazareff, B. et al. 1999, *ApJ*, 514, 121
- Crutcher, R. M., Hakobian, N., & Troland, T. H. 2009, *ApJ*, 692, 844
- Crutcher, R. M., Wandelt, B., Heiles, C., Falgarone, E., & Troland, T. H. 2010, *ApJ*, 725, 466
- Dahlem, M., Aalto, S., Klein, U. et al. 1990, *A&A*, 240, 237
- Dahlem, M., Lisenfeld, U., & Golla, G. 1995, *ApJ*, 444, 119
- Dahlem, M., Petr, M. G., Lehnert, M. D., Heckman, T. M., & Ehle, M. 1997, *A&A*, 320, 731
- Dalla Vecchia, C., & Schaye, J. 2008, *MNRAS*, 387, 1431
- Davis, L. J., & Greenstein, J. L. 1951, *ApJ*, 114, 206
- de Avillez, M. A., & Breitschwerdt, D. 2005, *A&A*, 436, 585
- Dobbs, C. L., & Price, D. J. 2008, *MNRAS*, 383, 497
- Dobler, G., Draine, B., & Finkenbeiner, D. P. 2009, *ApJ*, 699, 1374
- Draine, B. T., & Lazarian, A. 1998, *ApJ*, 494, L19
- Drzazga, R. 2008, M.Sc. Thesis, Jagiellonian University Kraków
- Drzazga, R. T., Chyży, K. T., Jurusik, W., & Wiórkiewicz, K. 2011, *AA*, 533, A22
- Dumas, G., Schinnerer, E., Tabatabaei, F. S. et al. 2011, *AJ*, 141, 41
- Dumke, M. 1997, PhD Thesis, University of Bonn

- Dumke, M., & Krause, M. 1998, in *The Local Bubble and Beyond*, ed. D. Breitschwerdt et al. (Berlin: Springer), 555
- Dumke, M., Krause, M., Wielebinski, R., & Klein, U. 1995, *A&A*, 302, 691
- Duncan, A. R., Haynes, R. F., Jones, K. L., & Stewart, R. T. 1995, *MNRAS*, 277, 36
- Duncan, A. R., Haynes, R. F., Jones, K. L., & Stewart, R. T. 1997, *MNRAS*, 291, 279
- Duncan, A. R., Reich, P., Reich, W., & Fürst, E. 1999, *A&A*, 350, 447
- Duric, N., & Seaquist, E. R. 1988, *ApJ*, 326, 574
- Ehle, M. 1995, Ph.D. Thesis, Univ. of Bonn
- Ehle, M., & Beck, R. 1993, *A&A*, 273, 45
- Ehle, M., Beck, R., Haynes, R. F. et al. 1996, *A&A*, 306, 73
- Ellis, R. S., & Axon, D. J. 1978, *Ap&SS*, 54, 425
- Elmouttie, M., Haynes, R. F., Jones, K. L. et al. 1995, *MNRAS*, 275, L53
- Elvius, A. 1962, *Bull Lowell Obs*, 5, 281
- Fabbiano, G., Klein, U., Trinchieri, G., & Wielebinski, R. 1987, *ApJ*, 312, 111
- Fendt, Ch., Beck, R., Lesch, H., & Neining, N. 1996, *A&A*, 308, 713
- Fendt, Ch., Beck, R., & Neining, N. 1998, *A&A*, 335, 123
- Fermi, E. 1949, *Phys Rev*, 75, 1169
- Ferrière, & K. 2009, *A&A*, 505, 1183
- Fiebig, D., & Güsten, R. 1989, *A&A*, 214, 333
- Fish, V. L., Reid, M. J., Argon, A. L., & Menten, K. M. 2003, *ApJ*, 596, 328
- Fletcher, A., Berkhuisen, E. M., Beck, R., & Shukurov, A. 2004, *A&A*, 414, 53
- Fletcher, A., Beck, R., Shukurov, A., Berkhuisen, E. M., & Horellou, C. 2011, *MNRAS*, 412, 2396
- Fosalba, P., Lazarian, A., Prunet, S., & Tauber, J. A. 2002, *ApJ*, 546, 762
- Fürst, E., Reich, W., Reich, P., & Reif, K. 1990, *A&AS*, 85, 691
- Gaensler, B. M., Dickey, J. M., McClure-Griffiths, N. M. et al. 2001, *ApJ*, 549, 959
- Gaensler, B. M., Haverkorn, M., Staveley-Smith, L. et al. 2005, *Science*, 307, 1610
- Gaensler, B. M., Beck, R., & Feretti, L. 2004, in *Science with the Square Kilometer Array*, *New Astronomy Reviews*, Vol. 48, ed. C. Carilli, & S. Rawlings (New York: Elsevier), 1003
- Gao, X. Y., Reich, W., Han, J. L. et al. 2010, *A&A*, 515, A64
- Georgelin, Y. M., & Georgelin, Y. P. 1976, *ApJ*, 49, 57
- Golla, G., & Hummel, E. 1994, *A&A*, 284, 777
- Gräve, R., & Beck, R. 1988, *A&A*, 192, 66
- Gräve, R., Klein, U., & Wielebinski, R. 1990, *A&A*, 238, 39
- Gray, A., Landecker, T. L., Dewdney, P. E., & Taylor, A. R. 1998, *Nature*, 393, 660
- Greaves, J. S., Holland, W. S., Jenness, T., & Hawarden, T. G. 2000, *Nature*, 404, 732
- Greenhill, L. J., Jiang, D. R., Moran, J. M. et al. 1995, *ApJ*, 440, 619
- Gressel, O., Elstner, D., Ziegler, U., & Rüdiger, G. 2008, *A&A*, 486, L35
- Guzmán, A. E., May, J., Alvarez, H., & Maeda, K. 2011, *A&A*, 525, A138
- Han, J. L. 2008, *IAUS*, 242, 55
- Han, J. L., & Zhang, J. S. 2007, *A&A*, 464, 609
- Han, J. L., Beck, R., & Berkhuisen, E. M. 1998, *A&A*, 335, 1117
- Han, J. L., Beck, R., Ehle, M., Haynes, R. F., & Wielebinski, R. 1999, *A&A*, 348, 405
- Han, J. L., Manchester, R. N., Lyne, A. G., Qiao, G. J., & van Straten, W. 2006, *ApJ*, 642, 868
- Han, J. L., Demorest, P. B., van Straten, W., & Lyne, A. G. 2009, *ApJS*, 181, 557
- Hanasz, M., Otmianowska-Mazur, K., Kowal, G., & Lesch, H. 2009, *A&A*, 498, 335
- Harnett, J. I., Haynes, R. F., Klein, U., & Wielebinski, R. 1989, *A&A*, 216, 39
- Harnett, J. I., Haynes, R. F., Wielebinski, R., & Klein, U. 1990, *Proc Astron Soc Aust*, 8, 257
- Harnett, J., Ehle, M., Fletcher, A. et al. 2004, *A&A*, 421, 571
- Haslam, C. G. T., Salter, C. J., Stoffel, H., & Wilson, W. E. 1982, *A&AS*, 47, 1
- Haverkorn, M., Katgert, P., & de Bruyn, A. G. 2003, *A&A*, 403, 1031, and 403, 1045
- Haverkorn, M., Katgert, P., & de Bruyn, A. G. 2004, *A&A*, 427, 169, and 427, 549
- Haverkorn, M., Gaensler, B. M., McClure-Griffiths, N. M. et al. 2006, *ApJS*, 167, 230
- Haynes, R. F., Klein, U., Wielebinski, R., & Murray, J. D. 1986, *A&A*, 159, 22
- Haynes, R. F., Klein, U., Wayte, S. R. et al. 1991, *A&A*, 252, 475
- Heald, G. 2009, in *Cosmic Magnetic Fields, from Planets to Stars and Galaxies*, ed. K. G. Strassmeier et al. (Cambridge: Cambridge University Press), 591
- Heald, G., Braun, R., & Edmonds, R. 2009, *A&A*, 503, 409
- Heesen, V., Beck, R., Krause, M., & Dettmar, R.-J. 2009a, *A&A*, 494, 563
- Heesen, V., Krause, M., Beck, R., & Dettmar, R.-J. 2009b, *A&A*, 506, 1123
- Heesen, V., Rau, U., Rupen, M. P., Brinks, E., & Hunter, D. A. 2011a, *ApJ*, 739, L23
- Heesen, V., Beck, R., Krause, M., & Dettmar, R.-J. 2011b, *A&A*, 535, A79
- Heiles, C. 2000, *AJ*, 119, 923
- Hildebrand, R. H., Kirby, L., Dotson, J. L. et al. 2009, *ApJ*, 696, 567
- Hiltner, W. A. 1958, *ApJ*, 128, 9

- Hinshaw, G., Weiland, J. L., Hill, R. S. et al. 2009, *ApJS*, 180, 225
- Hoang, T., & Lazarian, A. 2008, *MNRAS*, 388, 117
- Horellou, C., Beck, R., Berkhuijsen, E. M., Krause, M., & Klein, U. 1992, *A&A*, 265, 417
- Houde, M., Vaillancourt, J. E., Hildebrand, R. H., Chitsazzadeh, S., & Kirby, L. 2009, *ApJ*, 706, 1504
- Hummel, E., & Beck, R. 1995, *A&A*, 303, 691
- Hummel, E., Krause, M., & Lesch, H. 1989, *A&A*, 211, 266
- Hummel, E., Beck, R., & Dahlem, M. 1991, *A&A*, 248, 23
- Johnston-Hollitt, M., Hollitt, C. P., & Ekers, R. 2003, in *The Magnetized Interstellar Medium*, ed. B. Uyaniker et al. (Katlenburg: Copernicus), 13
- Jonas, J. L., Baart, E. E., & Nicolson, G. D. 1998, *MNRAS*, 297, 977
- Jones, T. J. 1989, *AJ*, 98, 2062
- Junkes, N., Fürst, E., & Reich, W. 1987, *A&A*, 69, 451
- Kallas, E., & Reich, W. 1980, *A&AS*, 42, 227
- Kepley, A. A., Mühle, S., Everett, J. et al. 2010, *ApJ*, 712, 536
- Kiepenheuer, K. O. 1950, *Phys Rev*, 79, 738
- Klein, U., Wielebinski, R., & Morsi, H. 1988, *A&A*, 190, 41
- Klein, U., Weiland, H., & Brinks, E. 1991, *A&A*, 246, 323
- Klein, U., Haynes, R. F., Wielebinski, R., & Meinert, D. 1993, *A&A*, 271, 402
- Klein, U., Hummel, E., Bomans, D. J., & Hopp, U. 1996, *A&A*, 313, 396
- Knapik, J., Soida, M., Dettmar, R.-J., Beck, R., & Urbanik, M. 2000, *A&A*, 362, 910
- Kogut, A., Dunkley, J., Bennett, C. L. et al. 2007, *ApJ*, 665, 355
- Krause, F., & Beck, R. 1998, *A&A*, 335, 789
- Krause, M. 1990, in *Galactic and Intergalactic Magnetic Fields*, ed. R. Beck et al. (Dordrecht: Kluwer), 187
- Krause, M. 2009, *Rev Mex AA*, 36, 25
- Krause, M., & Löhner, A. 2004, *A&A*, 420, 115
- Krause, M., Hummel, E., & Beck, R. 1989a, *A&A*, 217, 4
- Krause, M., Beck, R., & Hummel, E. 1989b, *A&A*, 217, 17
- Krause, M., Wielebinski, R., & Dumke, M. 2006, *A&A*, 448, 133
- Kronberg, P. P., Perry, J. J., & Zukowski, E. L. H. 1992, *ApJ*, 387, 528
- Kronberg, P. P., Bernet, M. L., Miniati, F. et al. 2008, *ApJ*, 676, 70
- Lacki, B. C., Thompson, T. A., & Quataert, E. 2010, *ApJ*, 717, 1
- Laine, S., & Beck, R. 2008, *ApJ*, 673, 128
- Landecker, T. L., & Wielebinski, R. 1970, *Aust J Phys Astrophys Suppl* 16, 1
- Landecker, T. L., Reich, W., Reid, R. I. et al. 2010, *A&A*, 520, A80
- Lazar, M., Schlickeiser, R., Wielebinski, R., & Poedts, S. 2009, *ApJ*, 693, 1133
- Lesch, H., Schlickeiser, R., & Crusius, A. 1988, *A&A*, 200, L9
- Levin, S. M., Langer, W. D., Kuiper, T. B. H. et al. 2000, *AAS*, 197, 1016
- Li, H., Dowell, C. D., Goodman, A., Hildebrand, R., & Novak, G. 2009, *ApJ*, 704, 891
- Lisenfeld, U., Völk, H. J., & Xu, C. 1996, *A&A*, 314, 745
- Manchester, R. N. 1974, *ApJ*, 188, 636
- Mao, S. A., Gaensler, B. M., Stanimirović, S. et al. 2008, *ApJ*, 688, 1029
- Mao, S. A., Gaensler, B. M., Haverkorn, M. et al. 2010, *ApJ*, 714, 1170
- Mathewson, D. S., & Milne, D. K. 1965, *Aust J Phys* 18, 635
- Mathewson, D. S., & Ford, V. L. 1970a, *Mem. RAS*, 74, 139
- Mathewson, D. S., & Ford, V. L. 1970b, *ApJ*, 160, L43
- Men, H., Ferrière, K., & Han, J. L. 2008, *A&A*, 486, 819
- Mitra, D., Wielebinski, R., Kramer, M., & Jessner, A. 2003, *A&A*, 398, 993
- Moss, D., & Shukurov, A. 1996, *MNRAS*, 279, 229
- Moss, D., Shukurov, A., & Sokoloff, D. et al. 2001, *A&A*, 380, 55
- Moss, D., Sokoloff, D., & Beck, R., Krause, M. 2010, *MNRAS*, 512, A61
- Mouschovias, T. Ch., & Tassis, K. 2009, *MNRAS*, 400, L15
- Murphy, E. 2009, *ApJ*, 706, 482
- Neininger, N. 1992, *A&A*, 263, 30
- Neininger, N., Klein, U., Beck, R., & Wielebinski, R. 1991, *Nature*, 352, 781
- Neininger, N., Beck, R., Sukumar, S., & Allen, R. J. 1993, *A&A*, 274, 687
- Neronov, A., & Vovk, I. 2010, *Science*, 328, 73
- Niklas, S. 1995, Ph.D. thesis, University of Bonn
- Niklas, S., & Beck, R. 1997, *A&A*, 320, 54
- Nord, M. E., Lazio, T. J. W., Kassim, N. E. et al. 2004, *ApJ*, 128, 1646
- Nota, T., & Katgert, P. 2010, *A&A*, 513, A65
- Noutsos, A., Johnston, S., Kramer, M., Karastergiou, A. 2008, *MNRAS*, 386, 1881
- Novak, G., Dotson, J. L., Dowell, C. D. et al. 2000, *ApJ*, 529, 241
- Otmianowska-Mazur, K., Elstner, D., Soida, M., & Urbanik, M. 2002, *A&A*, 384, 48
- Parker, E. N. 1979, *Cosmical Magnetic Fields* (Oxford: Clarendon Press)

- Patrikeev, I., Fletcher, A., Stepanov, R. et al. 2006, *A&A*, 458, 441
- Phillips, S., Kearsley, S., Osbourne, J. L. et al. 1981, *A&A*, 98, 286
- Prouza, M., & Šmida, R. 2003, *A&A*, 410, 1
- Rand, R. J., & Kulkarni, S. R. 1989, *ApJ*, 343, 760
- Reber, G. 1944, *ApJ*, 100, 279
- Rees, M. J. 2005, in *Cosmic Magnetic Fields*, ed. R. Wielebinski, & R. Beck (Berlin: Springer), 1
- Reich, W. 1982, *A&AS*, 48, 219
- Reich, W. 2003, *A&A*, 401, 1023
- Reich, W. 2006, *Cosmic Polarization*, ed. R. Fabri (Trivandrum: Research Signpost), 91
- Reich, W., & Reich, P. 1986, *A&AS*, 63, 205
- Reich, W., Fürst, E., Steffen, P. et al. 1984, *A&AS*, 58, 197
- Reich, W., Sofue, Y., Wielebinski, R., & Seiradakis, J. H. 1988, *A&A*, 191, 303
- Reich, W., Reich, P., & Fürst, E. 1990a, *A&AS*, 83, 539
- Reich, W., Fürst, E., Reich, P., & Reif, K. 1990b, *A&AS*, 85, 633
- Reich, P., Reich, W., & Fürst, E. 1997, *A&AS*, 126, 413
- Reich, P., Testori, J., & Reich, W. 2001, *A&A*, 376, 861
- Reich, W., Fürst, E., Reich, P. et al. 2004, in *The Magnetized Interstellar Medium*, ed. B. Uyanıker et al. (Katlenburg: Copernicus), 51
- Reif, K., Reich, W., Steffen, P. et al. 1987, *Mitt AG*, 70, 419
- Reuter, H.-P., Krause, M., Wielebinski, R., & Lesch, H. 1991, *A&A*, 248, 12
- Reuter, H.-P., Klein, U., Lesch, H., Wielebinski, R., & Kronberg, P. P. 1992, *A&A*, 256, 10
- Reuter, H.-P., Klein, U., Lesch, H., Wielebinski, R., & Kronberg, P. P. 1994, *A&A*, 282, 724
- Robishaw, T., Quataert, E., & Heiles, C. 2008, *ApJ*, 680, 981
- Roy, S., Prameh Rao, A., & Subrahmanyan, R. 2008, *A&A*, 478, 435
- Rudnick, L., & Brown, S. 2009, *AJ*, 137, 145
- Rüdiger, G., & Hollerbach, R. 2004, *The Magnetic Universe* (Weinheim: Wiley)
- Ruzmaikin, A. A., Shukurov, A. M., & Sokoloff, D. D. 1988, *Magnetic Fields of Galaxies* (Dordrecht: Kluwer)
- Scarrott, S. M., Ward-Thompson, D., & Warren-Smith, R. F. 1987, *MNRAS*, 224, 299
- Scarrott, S. M., Rolph, C. D., & Semple, D. P. 1990, in *Galactic and Intergalactic Magnetic Fields*, ed. R. Beck et al. (Dordrecht: Kluwer), 245
- Scarrott, S. M., Rolph, C. D., Wolstencroft, R. W., & Tadhunter, C. N. 1991, *MNRAS*, 249, 16P
- Schmidt, Th. 1976, *A&A Suppl* 24, 357
- Schnitzeler, D. H. F. M., Katgert, P., & de Bruyn, A. G. 2009, *A&A*, 494, 611
- Schoofs, S. 1992, Diploma Thesis, University of Bonn
- Segalovitz, A., Shane, W. W., & de Bruyn, A. G. 1976, *Nature*, 264, 222
- Seiradakis, J. H., Lasenby, A. N., Yusef-Zadeh, F. et al. 1985, *Nature*, 317, 697
- Shukurov, A. 2005, in *Cosmic Magnetic Fields*, ed. R. Wielebinski, & R. Beck (Berlin: Springer), 113
- Simard-Normandin, M., Kronberg, P. P. 1980, *ApJ*, 242, 74
- Soida, M. 2005, in *The Magnetized Plasma in Galaxy Evolution*, ed. K. T. Chyży et al. (Kraków: Jagiellonian University), 185
- Soida, M., Urbanik, M., & Beck, R. 1996, *A&A*, 312, 409
- Soida, M., Urbanik, M., Beck, R., & Wielebinski, R. 1999, *A&A*, 345, 461
- Soida, M., Urbanik, M., Beck, R., Wielebinski, R., & Balkowski, C. 2001, *A&A*, 378, 40
- Soida, M., Beck, R., Urbanik, M., & Braine, J. 2002, *A&A*, 394, 47
- Soida, M., Krause, M., Dettmar, R.-J., & Urbanik, M. unpubl., 531, *A&A*, A127
- Sokoloff, D., Shukurov, A., & Krause, M. 1992, *A&A*, 264, 396
- Sokoloff, D., Bykov, A. A., Shukurov, A. et al. 1998, *MNRAS*, 299, 189 and *MNRAS*, 303, 207 (Erratum)
- Stepanov, R., Frick, P., Shukurov, A., & Sokoloff, D. 2002, *A&A*, 391, 361
- Stepanov, R., Arshakian, T. G., Beck, R., Frick, P., & Krause, M. 2008, *A&A*, 480, 45
- Stil, J. M., Taylor, A. R., Dickey, J. M. et al. 2006, *AJ*, 132, 1158
- Stil, J. M., Krause, M., Beck, R., & Taylor, R. 2009, *ApJ*, 693, 1392
- Stix, M. 1975, *A&A*, 42, 85
- Sukumar, S., & Allen, R. J. 1989, *Nature*, 340, 537
- Sukumar, S., & Allen, R. J. 1991, *ApJ*, 382, 100
- Sun, X. H., & Reich, W. 2010, *Res Astron Astrophys*, 10, 1287
- Sun, X. H., Han, J. L., Reich, W. et al. 2007, *A&A*, 463, 993
- Sun, X. H., Reich, W., Waelkens, A., & Enßlin, T. A. 2008, *A&A*, 477, 573
- Sun, X. H., Reich, W., Han, J. L. et al. 2011, *A&A*, 527, A74
- Tabatabaei, F. S., Beck, R., Krügel, E. et al. 2007, *A&A*, 475, 133
- Tabatabaei, F. S., Krause, M., Fletcher, A., & Beck, R. 2008, *A&A*, 490, 1005
- Tang, Y.-W., Ho, P. T. P., Koch, P. M. et al. 2009, *ApJ*, 700, 251

- Taylor, A. R., Gibson, S. J., Peracaula, M. et al. 2003, *AJ*, 125, 3145
- Taylor, A. R., Stil, J. M., & Sunstrum, C. 2009, *ApJ*, 702, 1230
- Testori, J. C., Reich, P., & Reich, W. 2008, *A&A*, 484, 733
- Thompson, T. A., Quataert, E., Waxman, E., Murray, N., & Martin, C. L. 2006, *ApJ*, 645, 186
- Thum, C., & Morris, D. 1999, *A&A*, 344, 923
- Tüllmann, R., Dettmar, R.-J., Soida, M., Urbanik, M., & Rossa, J. 2000, *A&A*, 364, L36
- Tüllmann, R., Breitschwerdt, D., Rossa, J., Pietsch, W., & Dettmar, R.-J. 2006, *A&A*, 457, 779
- Urbanik, M., Elstner, D., & Beck, R. 1997, *A&A*, 326, 465
- Uyanıker, B., Fürst, E., Reich, W. et al. 1999, *A&AS*, 138, 31
- Vallée, J. P. 1996, *A&A*, 308, 433
- Vallée, J. P. 2005, *ApJ*, 619, 297
- van Albada, G. D., & van der Hulst, J. M. 1982, *A&A*, 115, 263
- Van Eck, C. L., Brown, J. C., Stil, J. M. et al. 2011, *ApJ*, 728, 97
- Verschuur, G. L. 1968, *Phys Rev Lett* 21, 775
- Vishniac, E. T., Lazarian, A., & Cho, J. 2003, in *Turbulence and Magnetic Fields in Astrophysics*, ed. E. Falgarone, & T. Passot (Berlin: Springer), 376
- Vollmer, B., Beck, R., Kenney, J. D. P., & van Gorkum, J. H. 2004, *AJ*, 127, 3375
- Vollmer, B., Soida, M., Beck, R. et al. 2007, *A&A*, 464, L37
- Vollmer, B., Soida, M., Chung, A. et al. 2010, *A&A*, 512, A36
- Weżgowiec, M., Urbanik, M., Vollmer, B. et al. 2007, *A&A*, 471, 93
- Weżgowiec, M., Urbanik, M., Vollmer, B. et al. 2012, *A&A* (in prep)
- Widrow, L. M. 2002, *Rev Mod Phys* 74, 775
- Wielebinski, R. 2006, *Astron Nachr* 327, 510
- Wielebinski, R., & Shakeshaft, J. R. 1964, *MNRAS*, 128, 19
- Wielebinski, R., Shakeshaft, J. R., & Pauliny-Toth, I. I. K. 1962, *Observatory*, 82, 158
- Wielebinski, R., Reich, W., Han, J. L., & Sun, X. H. 2008, *ASP Conf Ser*, 396, 13
- Wieringa, M., de Bruyn, A. G., Jansen, D. et al. 1993, *A&A*, 286, 215
- Wolleben, M., & Reich, W. 2004, *A&A*, 427, 537
- Wolleben, M., Landecker, T. L., Reich, W., & Wielebinski, R. 2006, *A&A*, 448, 411
- Wolleben, M., Landecker, T. L., Carretti, E. et al. 2009, in *Cosmic Magnetic Fields, from Planets, to Stars and Galaxies*, ed. K. G. Strassmeier et al. (Cambridge: Cambridge University Press), 89
- Wolleben, M., Landecker, T. L., Hovey, G. J. et al. 2010a, *AJ*, 139, 1681
- Wolleben, M., Fletcher, A., Landecker, T. L. et al. 2010b, *ApJ*, 724, L48
- Xiao, L., Han, J. L., Reich, W. et al. 2011, *A&A*, 529, A15
- Yates, K. W. 1968, *Aust J Phys* 21, 167
- Yusef-Zadeh, F., Morris, M., & Chance, D. 1984, *Nature*, 310, 557
- Yusef-Zadeh, F., Roberts, D. A., Goss, M. W. et al. 1996, *ApJ*, 466, L25

14 Astrophysics of Galactic Charged Cosmic Rays

Antonella Castellina¹ · Fiorenza Donato²

¹Osservatorio Astrofisico di Torino, Istituto Nazionale di Astrofisica, Torino, Italy

²Dipartimento di Fisica, Università di Torino, Torino, Italy

1	Introduction	727
1.1	Charged Cosmic Rays	729
1.2	Brief History: From Cosmic Rays to Astroparticle Physics	732
2	From 100 MeV n⁻¹ to 100 TeV n⁻¹	734
2.1	The Diffusion Equation	734
2.2	Experimental Methods for Direct Measurements	741
2.2.1	The Measure of Energy	742
2.2.2	The Measure of the Charge	743
2.3	Solutions to the Diffusive Equation and Comparison with Data	744
2.4	Antimatter in CRs	746
2.4.1	Antiprotons	747
2.4.2	Antideuterons	748
2.4.3	Positrons	750
3	From 100 TeV n⁻¹ to 100 PeV n⁻¹	752
3.1	Extensive Air Showers	752
3.1.1	The Electromagnetic Component	753
3.1.2	The Hadronic and Muonic Components	754
3.1.3	Cherenkov Light	755
3.1.4	Fluorescence Light	757
3.1.5	Radio Emission	757
3.2	Experimental Methods for Indirect Measurements	758
3.2.1	The Measure of the Charged Component	758
3.2.2	The Measure of the Cherenkov Light	759
3.2.3	The Measure of the Fluorescence	760
3.2.4	Energy and Composition Estimators	761
3.2.5	The Link to Particle Physics	764
3.3	The Knee Region	764
3.3.1	The Energy Spectrum	764
3.3.2	The Composition in the Knee Region	769
3.4	Models for the Knee	770

4	<i>Above 100 PeV n^{-1}: The Onset of Extragalactic Cosmic Rays</i>	773
4.1	Energy Spectrum and Composition	773
4.2	The Astrophysical Interpretation of the Transition	776
5	<i>The Measure of the Anisotropy</i>	779
5.1	Large Scale Anisotropy	780
5.2	Point Sources	782
6	<i>The Future of Cosmic-Ray Astrophysics</i>	783
	<i>Acknowledgments</i>	784
	<i>References</i>	785

Abstract: A review is given of the main properties of the charged component of galactic cosmic rays, particles detected at Earth with an energy spanning from tens of MeV up to about 10^{19} eV. After a short introduction to the topic and a historical overview, the properties of cosmic rays are discussed with respect to different energy ranges. The origin and the propagation of nuclei in the Galaxy are dealt with from a theoretical point of view. The mechanisms leading to the acceleration of nuclei by supernova remnants and to their subsequent diffusion through the inhomogeneities of the galactic magnetic field are discussed, and some clue is given on the predictions and observations of fluxes of antimatter, both from astrophysical sources and from dark matter annihilation in the galactic halo.

The experimental techniques and instrumentations employed for the detection of cosmic rays at Earth are described. Direct methods are viable up to $\approx 10^{14}$ eV, by means of experiments flown on balloons or satellites, while above that energy, due to their very low flux, cosmic rays can be studied only indirectly by exploiting the particle cascades they produce in the atmosphere.

The possible physical interpretation of the peculiar features observed in the energy spectrum of galactic cosmic rays, and in particular the so-called “knee” at about 4×10^{15} eV, is discussed. A section is devoted to the region between about 10^{18} and 10^{19} eV, which is believed to host the transition between galactic and extragalactic cosmic rays. The conclusion gives some perspectives on the cosmic ray astrophysics field. Thanks to a wealth of different experiments, this research area is living a very flourishing era. The activity is exciting both from the theoretical and the instrumental sides, and its interconnection with astronomy, astrophysics, and particle physics experiences nonstop growth.

Keywords: Cosmic Rays – Origin and Propagation – Instrumentation: Detectors – Energy spectrum and Composition – Galactic to extragalactic transition

1 Introduction

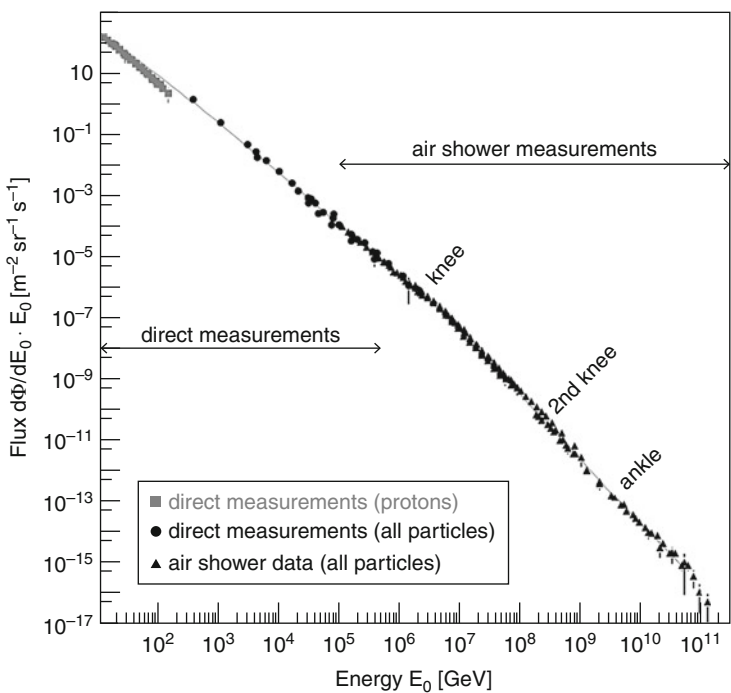
“The subject of cosmic rays is unique in modern physics for the minuteness of the phenomena, the delicacy of the observations, the adventurous excursions of the observers, the subtlety of the analysis, the grandeur of the inferences. (K.K. Darrow)”

A large contribution to the knowledge and understanding of the Galaxy is given by the observation of the most energetic particles, the cosmic rays (CRs). These relativistic particles, reaching the Earth from the outer space, are either primary nuclei, arriving directly from the sources, or secondary products of the spallation processes (i.e., fragmentation by nuclear destruction) taking place during the propagation from the sources through the interstellar medium (ISM).

Cosmic rays play an important role in the dynamics of the Galaxy: their energy density $\rho_E \approx 1 \text{ eV cm}^{-3}$ is comparable to the energy density of the visible starlight $\rho_S \approx 0.3 \text{ eV cm}^{-3}$, the galactic magnetic fields $B^2/2\mu_0 \approx 0.25 \text{ eV cm}^{-3}$ (if $B \approx 3 \mu\text{G}$) or the cosmic microwave background (CMB) radiation $\rho_{\text{CMB}} \approx 0.25 \text{ eV cm}^{-3}$. Such nonthermal component is indeed strictly linked to radiation and magnetic fields. Measuring cosmic rays and their properties could be an effective tool to understand the stellar nucleosynthesis and the Supernova evolution. These particles strongly influence the galactic chemical composition and evolution, and their interactions with the cosmic radiation background, the interstellar radiation field, and interstellar

gas give rise to diffuse gamma ray emission. The challenge we are facing is that of identifying the sources of cosmic rays and the mechanism through which low-energy particles (the seeds being just single elements, or dust and grains) are accelerated to such high energy to be called cosmic rays and of understanding their propagation through the galactic magnetic fields.

The cosmic-ray extreme energies are by far beyond the reach of the most powerful man-made accelerators, so that they are of great interest also from the point of view of particle physics, probing the standard models of hadronic interactions and the laws of relativity in extreme domains.

The flux of cosmic rays is seen at Earth varying about 32 orders of magnitude across an energy range spanning 14 orders of magnitude. It amounts to $\approx 10^4 \text{ m}^{-2} \text{ s}^{-1}$ at $\approx 10^6 \text{ eV}$ to less than $1 \text{ km}^{-2} \text{ century}$ at $\approx 10^{19} \text{ eV}$. The all-particle energy spectrum is shown in , where the main structures of an otherwise almost pure power law are visible. Changes of slope are taking place at the *knee*, at $\approx 3\text{--}4 \text{ PeV}$,¹ the *second knee*, near 400 PeV , and the *ankle*, a broader feature around 3 EeV . The investigation of these features is a valuable tool for the study of the cosmic rays, their nuclear composition, and the energy above which an extragalactic component takes over.

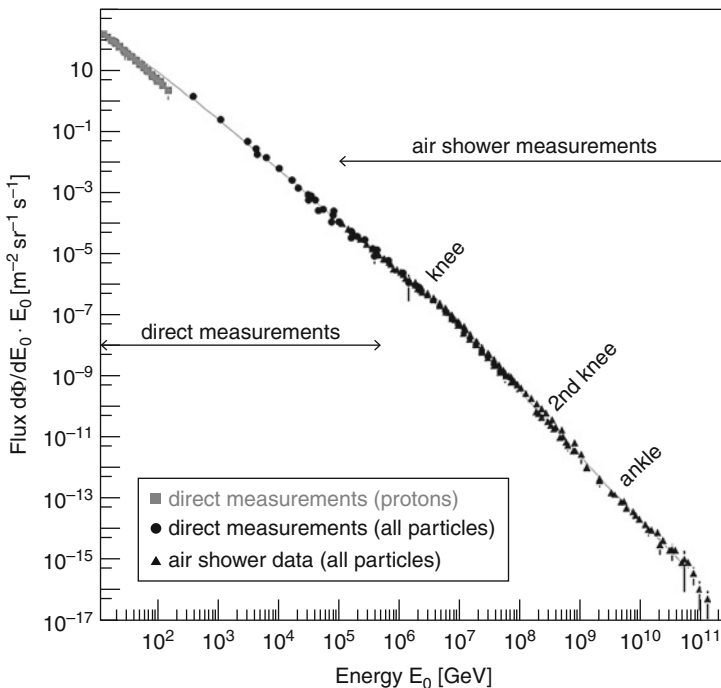


 Fig. 14-1

The all-particle energy spectrum. The changes of slope correspond to the indicated knee and ankle energies (Modified from Blümer et al. (2009))

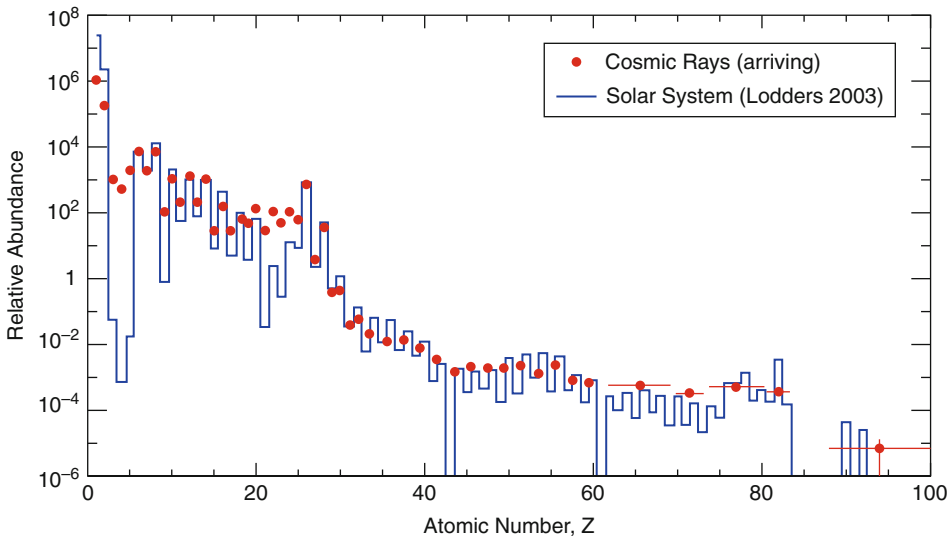
¹Energy is hereafter measured in units of 1 GeV, 1 TeV, 1 PeV, and 1 EeV meaning 10^9 , 10^{12} , 10^{15} , and 10^{18} eV.

Due to the wide energy range and the rapidly changing flux, it is obvious that different experimental techniques are needed in diverse energy regions. Direct measurements of the primary cosmic rays can be performed up to 10^{14} eV. Above this threshold, the low fluxes due to the steeply falling spectrum force us to exploit indirect methods, detecting the extensive air showers generated by the interaction of cosmic rays in the atmosphere.

1.1 Charged Cosmic Rays

Galactic cosmic rays (GCR) are supposed to be produced by nucleosynthesis processes in the stellar interior, where they are fractionated, ejected and then accelerated to CR energies, and diffused throughout the whole Galaxy. When reaching the solar system – where the slowest CRs experience a spectral distortion by the solar wind – they can be observed both by space-based experiments or by detectors at ground.

Starting from the spectrum and composition of the cosmic rays observed at Earth, and knowing the characteristics of propagation of these particles in the Galaxy, which include diffusion and convection, energy losses, reacceleration, and production of secondaries through spallation processes, it is possible to step back from the observed spectrum at Earth to the source one. The similarity between the cosmic-ray composition as measured directly with detectors on satellites or balloons, like BESS (Wang et al. 2002), HEAO-3 (Engelmann et al. 1990), and ACE/CRIS (Stone et al. 1998; Wiedenbeck et al. 2007), and that found in the solar system (Lodders 2003) is shown in [Fig. 14-2](#). The overabundance of elements like Li, Be, B ($Z = 3/5$), and sub-Fe ($Z = 21/25$) in cosmic rays can be explained by their origin: they are in fact secondaries produced in spallation processes during propagation in the interstellar medium.



■ Fig. 14-2

Elemental abundances in the arriving cosmic rays for $E \leq 1 \text{ GeV n}^{-1}$ near the peak of their spectrum (relative to carbon) compared to those found in the solar system material (Wiedenbeck et al. 2007)

The problem of the origin of cosmic rays is related to many different issues: What are the seeds out of which the sources accelerate particles, what accelerators have the power to boost them to such high energies, and which part of the cosmic-ray spectrum is of galactic origin? Charged particles are bent in the galactic magnetic fields, so there is no obvious way of tracking them back to their sources. The energy spectrum which is observed at Earth is folded with the source one through an energy-dependent diffusion coefficient which shapes the effects of the galactic magnetic field in an effective way and which can be deduced only phenomenologically. These peculiarities are at variance with γ rays, which point more directly to the sources: the accelerated ions around a source are expected to interact with the ambient matter and produce high-energy γ rays which can be detected at Earth.

On the basis of energetic arguments, supernova remnants (SNR) have long been considered the most probable sources of galactic cosmic rays (Ginzburg and Syrovatskii 1964). In fact, the average kinetic energy of a SN ejecta is $L_{\text{kin}} \simeq 1.6 \times 10^{51}$ erg (for ten solar masses traveling at $\simeq 4,000$ km s $^{-1}$). Assuming a supernova rate of explosion in our Galaxy of the order of 30 year $^{-1}$ and an efficiency for converting the kinetic energy into relativistic particles of $\simeq 10\%$, SNRs can provide the $\simeq 10^{41}$ erg s $^{-1}$ power required to keep the cosmic-ray energy density. The energetic argument is however not conclusive, since other potential sources could meet the requirement, like stars with powerful winds or pulsars.

Based on the first theory of cosmic-ray acceleration developed by Fermi (1949), the most accredited mechanism to convert from the kinetic motion of the plasma to kinetic energy of charged particles is the diffusive acceleration in presence of shock waves (DSA) powered by supernova explosions propagating from the remnant to the interstellar medium (Berezhko et al. 1999). Traversing the boundary between the unshocked upstream and the shocked downstream region back and forth, charged particles gain each time an energy $\Delta E \propto E$; the acceleration spectrum follows a power law in momentum, $Q(E) \propto p^{-\alpha}$, with α located between 2.0 and 2.5. Taking into account the diffusion of cosmic rays in the Galaxy, with a diffusion coefficient K expected to be proportional to the rigidity of the particle ($R = pc/Ze$, where Z is the charge and c is the speed of light) as $K \propto R^\delta$, this will eventually lead to the observed spectrum at Earth $N_{\text{obs}}(E) \propto E^{-\gamma}$, with $\gamma = \alpha + \delta$, up to a maximum energy:

$$E_{\text{max}} \simeq Ze(B/\mu\text{G})(L/pc)\beta_{\text{shock}} \text{PeV}, \quad (14.1)$$

where Ze is the particle charge, B is the galactic magnetic field strength, L is the linear dimension of the acceleration site, and β_{shock} is the shock velocity (Hillas 1984).

Direct measurements of the cosmic rays performed by means of balloons or satellites can help in the study of both diffusion and acceleration spectrum. The experimental measurement of the secondary-to-primary ratio, that is, the boron-to-carbon flux ratio B/C , gives the most direct information on the slope of the diffusion coefficient, δ (see discussion in [Sect. 2.1](#)). A careful measurement of the fluxes of individual elements can determine the spectrum slope γ for light nuclei with a negligible error, bringing information on the acceleration slope power index α after having taken into account the uncertainty due to δ .

Besides the measurement of the all-particle energy spectrum, remarkable achievements of extensive air shower detectors include the determination of the energy spectra of single or groups of elements, showing cutoff energies at constant rigidity E/Z , even if spectral steepenings at constant energy/nucleon E/A , predicted by some of the models, cannot yet be disproven.

The interpretation of the knee feature in the energy spectrum is another important clue to understand the origin of the galactic cosmic rays. Different models attribute the knee either to



a limit on the maximum energy attainable during acceleration in the source or to the leakage of particles out of the Galaxy during their propagation. Other ideas point to interactions with background particles like massive neutrinos or photodisintegration of nuclei in the fields of soft photons, or finally to new properties of hadronic interactions taking over at high energy.

The region between about 10^{18} and 10^{19} eV is of particular interest, as it is supposed to host the transition from galactic to extragalactic cosmic rays (EGCR): cosmic-ray nuclei of charge Z and energy E cannot be confined in the Galaxy if their Larmor radius $r_L \simeq E_{\text{PeV}} / (Z B_{\mu\text{G}})$ pc becomes larger than the transversal dimension of the galactic disk ($\simeq 100$ pc).

For a proton, the confinement becomes impossible above about 1 EeV; a factor 26 higher energy would be required for iron.


The transition is described by the current models in terms of source emissivity and energy spectrum of particles at the source; the source density n is assumed to be a function of the red shift z , with a source evolution parameter m such that $dn/dz \propto (1+z)^m$.

In the “*dip*” model (Berezinsky et al. 2006), the transition is due to e^+e^- pair production by extragalactic protons on the photons of the cosmic microwave background and takes place at the second knee. Only a very low contamination ($\leq 10/15\%$) by heavier nuclei can be allowed in an otherwise pure proton beam. The “*mixed*” model (Globus et al. 2008), on the contrary, puts the transition around $\simeq 3$ EeV; the nuclear composition is mixed as in the galactic component, and the dip is filled by the contribution of elements other than protons. The more traditional “*ankle*” model (Hillas 2005) considers the ankle as the intersection between a very steep galactic component and a flat extragalactic one. Composition and anisotropy of high-energy cosmic rays are the most useful tools to discriminate among the models.

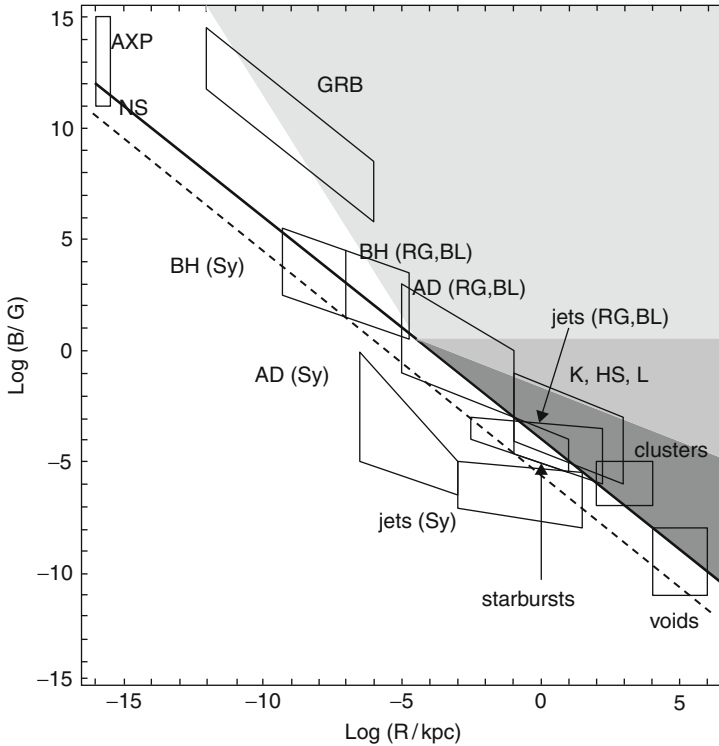
Where is the natural end of the cosmic-ray spectrum? As shown in  Fig. 14-3, only very few astrophysical objects are thought to be able to accelerate particles to ultra high energies (UHE), above \simeq few 10^{19} eV : neutron stars (NS), gamma ray bursts (GRB), and radio Galaxies. The acceleration sites in radio Galaxies (indicated as Seyfert-Sy, blazars-BL, radio-loud FR-I and FR-II) include both the region close to the black hole (BH) and the accretion disk (AD) and extended structures like knots (K) and hot spots (HS) in jets and lobes (L). The shaded areas are allowed by the radiation-loss constraints (shown for protons only). A given energy can be obtained either in a large region with low magnetic field or in a compact high field source (see  Eq. 14.1).

As recognized soon after the discovery of the cosmic microwave background, cosmic rays of ultrahigh energy undergo interactions with the intergalactic radiation fields, giving rise to an attenuation of their flux, the well known GZK effect (Greisen 1966; Zatsepin and Kuzmin 1966).

For protons, the most important interactions with the microwave background are the pion production ($p + \gamma_{\text{CMB}} \rightarrow N + \pi$) and the production of e^+e^- pairs ($p + \gamma_{\text{CMB}} \rightarrow p + e^+ + e^-$). The energy thresholds for these two processes are respectively given by $\simeq 6 \times 10^{19}$ and $\simeq 4 \times 10^{17}$ eV, but the energy loss per interaction for pion production is much higher (20% compared to 0.1% for the pair production). For heavier nuclei, the most important energy losses are due to photodisintegration ($A + \gamma_{\text{CMB}} \rightarrow (A-1) + N$) and pair production ($A + \gamma_{\text{CMB}} \rightarrow A + e^+ + e^-$) on the photons of the cosmic microwave background.

The fraction of cosmic rays that arrive at Earth from a given distance for nuclei above different energies is shown in  Fig. 14-4, where the GZK horizon is defined as the distance from which 50% of primaries originate.

A discussion of the extragalactic spectrum features, above the transition, is outside the scope of this review. It is however important to remind that at these extreme energies, a clear evidence



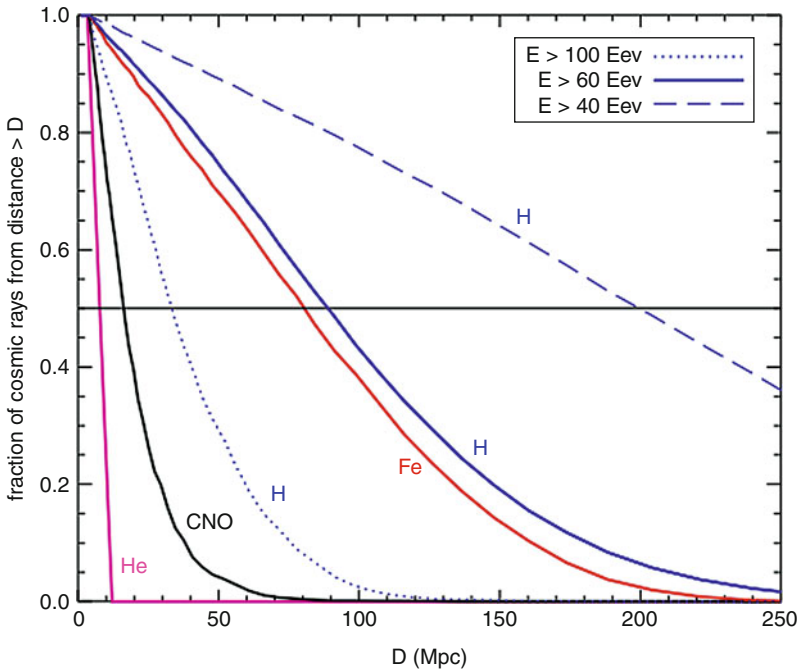
■ Fig. 14-3

The magnetic field and size of potential cosmic-ray accelerators (Ptitsyna and Troitsky 2010). *Full and dashed lines*: the lower boundary allowed by the Hillas criterium for 10^{20} eV protons and iron nuclei

for a flux suppression has been measured both by the HiRes (Abbasi et al. 2008) and Pierre Auger (Abraham et al. 2008) collaborations, consistent with that expected due to the GZK effect. However, the same effect could also be related to the maximum energy of sources. The topic is discussed at length in the reviews of Nagano (2009) and Kotera and Olinto (2011).

1.2 Brief History: From Cosmic Rays to Astroparticle Physics

About a century ago, cosmic rays were identified as being a source of radiation on Earth. The proof came from two independent experiments. The Italian physicist Domenico Pacini observed the radiation strength to decrease when going from the ground to few meters underwater (both in a lake and in the sea) (Pacini 1912). Victor Hess during several ascents with hydrogen-filled balloons up to altitudes of 5 km (Hess 1912) measured the ionization rate of air as a function of altitude. He explained the measured increase of ionizing radiation with increasing height with the presence of a radiation continuously penetrating into the Earth atmosphere from outer space; for this discovery, he was awarded the Nobel Prize in 1936. Hess measurements were further extended up to 9 km by the subsequent ascents of Kolhorster (1925).



■ Fig. 14-4

GZK horizons of p , He, CNO, and Fe above different energies (Kotera and Olinto 2011). Light nuclei are promptly dissociated by CMB photons, while p and Fe may reach us from distances $\lesssim 100$ Mpc

The first hints for a corpuscular nature of the cosmic radiation came exploiting the newly invented Geiger-Muller counters in the experiment of Bothe and Kolhorster (1929) and from measurements by J. Clay, who showed that the intensity of cosmic rays depends on the magnetic latitude of the observer (Clay 1930) and they thus had to be charged particles.

Using a cloud chamber, in 1927, D. Skobelzyn observed the first tracks left by cosmic rays. In 1934, Bruno Rossi reported an observation of near-simultaneous discharges of two Geiger counters widely separated in a horizontal plane, noting that “it seems that extended showers of particles every now and then reach our detectors...determining temporal coincidences among counters even quite far away from each other...” By 1938, using Wilson chambers and Geiger tubes spread over a wide area and working in coincidence, Pierre Auger showed that very high-energy cosmic rays trigger extensive air showers in the Earth atmosphere, sharing the primary energy among billions of lower energy particles that reach the ground together (Auger et al. 1938; Kolhorster 1938). On the basis of his measurements, Auger concluded that he had observed showers with energies of 10^{15} eV, incredibly higher compared to any previously measured before.

From the 1930s to the 1950s, before man-made accelerators reached very high energies, cosmic rays served as a source of particles for high-energy physics investigations and led to the discovery of a wealth of subatomic particles, such as the muon, the pion, hyperons, etc. The field benefited by many different technical improvements, from Wilson chambers to Geiger-Muller tubes to photographic emulsions. The development of photomultipliers allowed the use

of scintillators and Cherenkov detectors. Experimental measurements were performed in very different environments, with balloons in the highest atmosphere, high altitude observatories, and more recently underground laboratories.

Since the mid-1940s large apparatuses were used to detect the extensive air showers. In 1958, G.V. Kulikov and G.B. Khristiansen measured the integral electron number spectrum in air showers using an array of hodoscope counters (Kulikov and Khristiansen 1958); this brought to the first detection of the “knee” around few PeV.

In 1962, using an air shower array in Volcano Ranch, New Mexico, an event with an energy of tens of Joules (about 10^{20} eV) was observed (Linsley 1963). In 1966, an abrupt steepening of the cosmic-ray spectrum above 10^{20} eV was predicted (Greisen 1966; Zatsepin and Kuzmin 1966) as a result of the interactions of the cosmic rays with the newly discovered cosmic microwave background radiation, since then called the “GZK cutoff.” In the subsequent years, the ultrahigh-energy events have been studied either by means of scintillator or water detectors in ground arrays at Haverah Park (Edge et al. 1973), SUGAR (Bell 1974), Yakutsk (Afanasiev 1993), Akeno (Nagano et al. 1992a), and AGASA (Chiba et al. 1992) or by means of a new technique exploiting air fluorescence in atmosphere at Fly’s Eye (Baltrusaitis et al. 1985) and HiRes (Bird et al. 1993). Most recently, both techniques have been employed in the Pierre Auger Observatory (Abraham et al. 2004).

Since 1927, when R. Millikan introduced the term “cosmic rays,” the main focus of this research has been directed towards the astrophysical investigations of their origin, acceleration and propagation, what role they play in the dynamics of the Galaxy, and what their composition tells us about matter from outside the solar system. Starting at the end of the 1980s, the new interdisciplinary field of “astroparticle physics” was born, in which astrophysics, cosmology, cosmic rays, and particle physics together contribute to shed light on the nature and structure of the matter in the universe.

2 From 100 MeV n⁻¹ to 100 TeV n⁻¹

2.1 The Diffusion Equation

The spectrum of galactic cosmic rays – spanning from tens of MeV/nucleon(n) up to $\gtrsim 10^{18}$ eV – is fundamentally shaped by acceleration and diffusion, at least for energies $\gtrsim 10^2 - 10^3$ GeV n⁻¹. Several other phenomena – for instance, convection, reacceleration, nuclear fragmentation, electromagnetic losses, and solar modulation – compete at lower energies, where their effect is often degenerate and prevents unambiguous interpretation of the wealth of data collected in the lower tail of the galactic spectrum. Nevertheless, phenomenological models able to reproduce data on a wide energy range can be built (the milestone in the field literature being Berezhinskii et al. (1990)). The most realistic propagation models are the diffusion ones, even if the so-called leaky box model has been often preferred in the past for its simplicity. In the leaky box model, the densities of sources q^j , of interstellar matter n , and of cosmic rays N^j are assumed to be homogeneous in a finite propagation volume delimited by a surface. In addition, each nucleus is supposed to escape from the leaky box with a probability per unit of time $1/\tau_{\text{esc}}$. In a steady-state regime, the relevant densities, for a given nucleus j , obey the following equation:

$$\frac{N^j}{\tau_{\text{esc}}} + nv\sigma^j N^j = q^j + \sum_{\text{heavier } k} nv\sigma^{kj} N^k, \quad (14.2)$$

where v is the nucleus velocity and σ^j is its destruction cross section. The leaky box model has been successful to explain most of the observed fluxes for stable species by the single function $\tau_{\text{esc}}(E)$. This function is usually adjusted to the data, its physical interpretation being found afterwards, or extracted directly from more complete propagation equations (Jones et al. 2001).

At variance, diffusive models account for spatial dependence of sources, CR densities, and in principle of the interstellar matter. The Galaxy is shaped as a thin gaseous disk where all the astrophysical sources are located, embedded in a thick diffusive magnetic halo. Diffusive models, besides being more realistic and closer to a physical interpretation for each component, have proven to be successful in reproducing the nuclear, antiproton, and radioactive isotopes data. They also allow to treat contributions from dark matter (or other exotic) sources located in the diffusive halo.

The relevant transport equation for a charged particle wandering through the magnetic inhomogeneities of the galactic magnetic field writes in terms of the differential density $N(E, \vec{r})$ as a function of the total energy E and the position \vec{r} in the Galaxy. Assuming steady state ($\partial N/\partial t = 0$), the transport equation for a given nucleus (the subscript j is omitted) can be written in a compact form as

$$(-\vec{\nabla} \cdot (K\vec{\nabla}) + \vec{\nabla} \cdot \vec{V}_C + \Gamma_{\text{rad}} + \Gamma_{\text{inel}})N + \frac{\partial}{\partial E} \left(bN - c \frac{\partial N}{\partial E} \right) = \mathcal{S}. \quad (14.3)$$

The first bracket in the l.h.s. accounts for (1) spatial diffusion $K(\vec{r}, E)$, (2) convection with speed $\vec{V}_C(\vec{r})$, (3) the (possible, for some isotopes with half lifetime τ_0) radioactive decay rate $\Gamma_{\text{rad}}(E) = 1/(\gamma\tau_0)$ (γ here is the Lorentz factor), and (4) the destruction rate $\Gamma_{\text{inel}}(\vec{r}, E) = \sum_{\text{ISM}} n_{\text{ISM}}(\vec{r})v\sigma_{\text{inel}}(E)$ due to collisions with the interstellar medium. In this last expression, $n_{\text{ISM}}(\vec{r})$ is the density of the interstellar medium in the various locations of the Galaxy and in its different H and He components and $\sigma_{\text{inel}}(E)$ is the destruction (inelastic) cross section for a given nucleus. The coefficients b and c are, respectively first- and second-order gains/losses in energy. The source term \mathcal{S} includes primary sources of CRs (e.g., supernovae), secondary sources due to the fragmentation of heavier nuclei, and secondary decay-induced sources.




In the following, each of the components of (► Eq. 14.3) is reviewed.


Acceleration. As we already mentioned, the acceleration spectrum of primary galactic CRs is believed to be determined by supernovae (SN). The acceleration process takes place in the SN remnant (SNR) phase, and the most plausible mechanism is diffusive shock acceleration. Indeed, if the acceleration would take place in the explosion itself, the adiabatic losses suffered by charged particles would require an enormous energy amount to reach the observed energy levels (Drury et al. 1982, 2001). The spectrum of accelerated particles escaping the remnant follows a power law in momentum, $Q(E) \propto p^{-\alpha}$, with α located between 2.0 and 2.5. Even if a precise value for α cannot be predicted, the effective spectral index as derived by several indirect observational tests and by numerical studies is close to $\alpha \sim 2.0\text{--}2.1$ (Drury et al. 2001 and references therein; Berezhko et al. 1994).

The diffusive SNR shock acceleration model has been improved by the nonlinear reaction effects of the accelerated particles on the shock structure (Berezhko and Völk 2000; Meyer et al. 1997a). A complete review on the topic can be found in Malkov and Drury (2001).

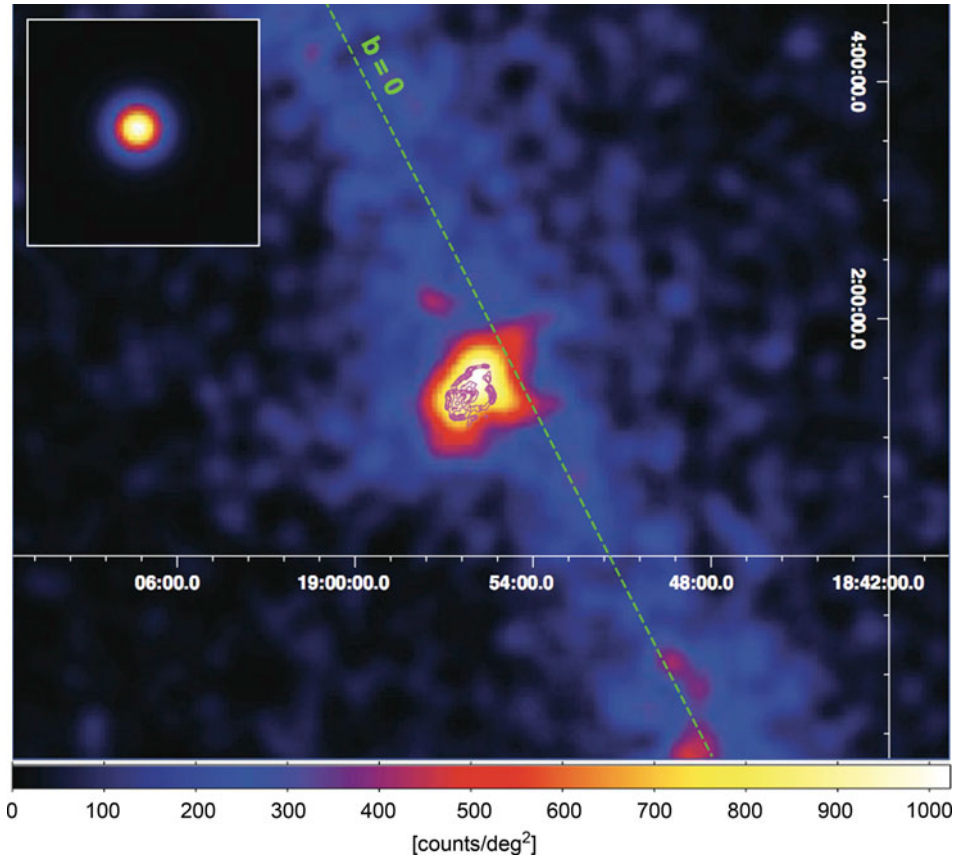
The structure of the shock is modified by the accelerated particles pressure acting on the background plasma, though in a collisionless manner. Particles with different momenta experience different effective accelerations. One consequence of this effect is that the spectrum *inside* the SNR is no more a pure power law but is concave as a function of momentum (Amato and Blasi 2006). Additionally, it was noticed in Lagage and Cesarsky (1983a, b) that the maximum

energy reachable by particles in a SNR is by far lower than the knee energy. A possible way out is to assume Bohm-like diffusion coefficient – which scales as the inverse of the magnetic field instead of the inverse squared field – and a strongly nonlinear magnetic field amplification. The latter is due to the reaction effects of the accelerated particles on the shock structure. In this scheme, the maximum momentum a proton can acquire is about few 10^6 GeV, thus satisfying the knee energy reach (Blasi et al. 2007).

A convincing evidence of particle acceleration in young SNRs would be the observation of high-energy γ s with a clear hadronic origin: due to their hard energy spectra, the relativistic particles produced at the source with high acceleration efficiency can interact with the interstellar medium, producing neutral and charged pions which in turn generate photons and neutrinos (Berezhko and Völk 2000; Drury et al. 1994). This signal must however be tagged against a competing emission process for γ s from inverse Compton scattering by high-energy electrons. To get an unambiguous determination of the origin of the radiation, γ -ray telescopes must provide spatially resolved spectral measurements and correlate their studies with other wavelength data (a deeper discussion can be found in the contribution by F. Aharonian, this book). As an example, RX J1713.7–3946 (Abdo et al. 2011; Aharonian et al. 2004, 2006; Enomoto et al. 2002; Zirakashvili and Aharonian 2010) is a source of TeV γ emission corresponding in morphology to the nonthermal X-ray emission seen by the ASCA satellite (Uchiyama et al. 2003) and to the CO observations by the NANTEN sub-mm telescope (Fukui et al. 2003). Similarly, gamma observations of the SNR W44 by the Fermi Large Area Telescope are in remarkable agreement with the morphology of the remnant in the radio image in the 20-cm wavelength, as displayed in  Fig. 14-5. The contribution from inverse Compton scattering of electrons cannot yet be excluded, but it seems unlikely that it could dominate in the GeV band. These results are visible in  Fig. 14-6, where contributions from π^0 decay are displayed along with electron bremsstrahlung, inverse Compton scattering, and bremsstrahlung from secondary electrons and positrons. A different treatment should be deserved for the acceleration and propagation of UHECRs, which are commonly believed to be of extragalactic origin (Blasi 2005; Hillas 1984 and references therein). This topic is discussed at length in  Sect. 4.

Source composition. The composition of CRs at their source can be inferred either as the solar system composition convoluted with the first ionization potential (FIP) (Casse and Goret 1978), or with the volatility (Meyer et al. 1997b), a parameter indicating the condensation temperature of an element (volatile elements are the light ones, while refractory elements are mostly metals, having high melting points). In the first case, it has been suggested that the seed material forming the cosmic rays be in coronal mass emissions from Sun-like stars, where low FIP elements are overabundant. Most interestingly, in the latter model, the seeds are instead refractory elements condensed in grains, which are accelerated to relativistic energies (thus, becoming “cosmic rays”) in supernova shocks with an efficiency proportional to their high charge-to-mass ratio.  Figure 14-7, taken from Wiedenbeck et al. (2007), shows the CR abundances in comparison with the solar system ones (Anders and Grevesse 1989). The first ones have been obtained using a leaky box model for interstellar propagation with a solar modulation one and normalized to the energy spectra of individual isotopes from CRIS data (Wiedenbeck et al. 2007). These elements come from different nucleosynthesis processes and stars with various initial masses: the striking similarity (within 20% taking into account the systematics) between the two populations can be understood only if the two samples are coming from a similar, well-mixed material.

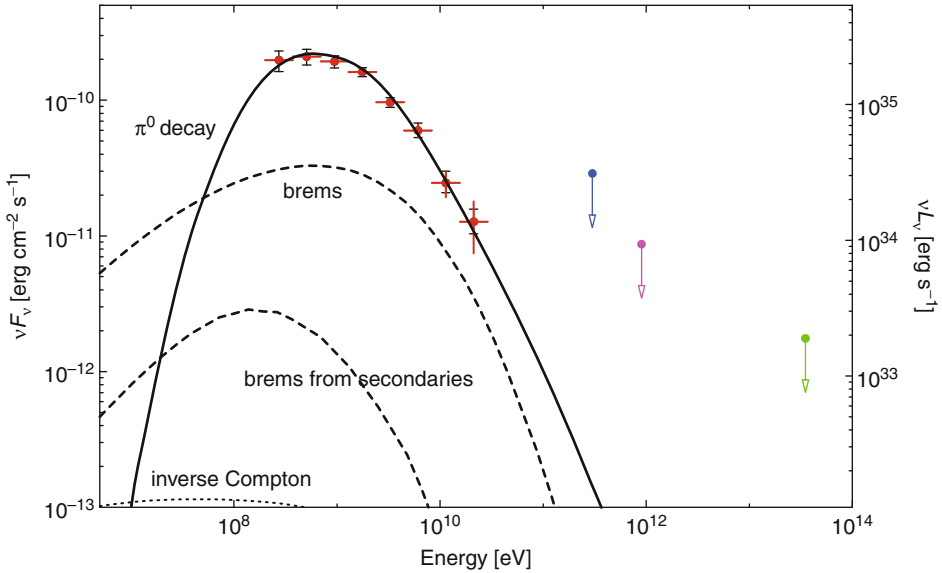
The largest differences are instead observed for the $^{22}\text{Ne}/^{20}\text{Ne}$ and at a lower extent for $^{58}\text{Fe}/^{56}\text{Fe}$. The isotopic composition of Ne, Fe, and other species has been measured by the



■ Fig. 14-5

Fermi-LAT image (2–10 GeV) of the region where the SNR W44 is located. The *color scale* indicates counts per solid angle on a linear scale. The *green line* corresponds to the galactic plane. The radio image of W44 as seen in 20-cm wavelength by the very large array VLA is overlaid as the magenta contour (Abdo et al. 2010a)

CRIS spectrometer (Wiedenbeck et al. 2007). The ratio $^{22}\text{Ne}/^{20}\text{Ne}$ so found is consistent with galactic cosmic ray sources with $\approx 80\%$ solar system composition plus a $\approx 20\%$ of material from Wolf-Rayet (WR) stars. These kinds of stars are evolutionary products of the hot and massive OB stars, which are short living and highly radiating objects, loosely organized in groups called OB associations. As a consequence, any model of origin of galactic cosmic rays must include an efficient mechanism to inject the WR material in the accelerator. Measurements of the ^{59}Ni , an isotope which decays only by electron capture, have shown that it has almost completely decayed to ^{59}Co , allowing to set up a limit of $\approx 10^5$ year for the time that has to elapse from the production of the material and its acceleration to cosmic-ray energies (Higdon and Lingenfelter 1999). Models of superbubbles environments (Higdon et al. 1998) host SN events every 0.3–3.5 My, thus allowing enough time for Ni to decay. With 50 days of data, the TIGER experiment (Link et al. 2009) measured a clear ordering of abundances of galactic cosmic rays when compared



■ Fig. 14-6

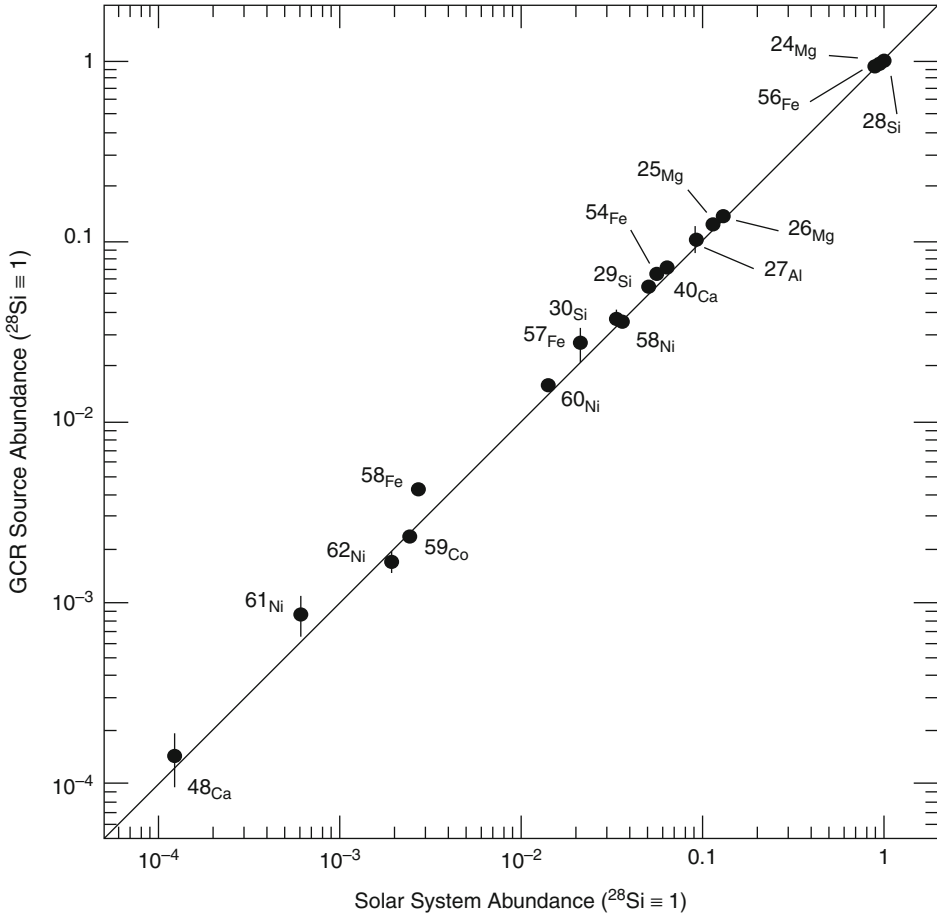
Fermi-LAT spectral energy distributions of SNR W44. Each curve corresponds to contributions from π^0 decay (solid), electron bremsstrahlung (dashed), inverse Compton scattering (dots), and bremsstrahlung from secondary electrons and positrons (Abdo et al. 2010a). Colored arrows refer to upper limits from Whipple, Heger, and Milagro experiments

to a mixture of 80% solar system and 20% star outflow, thus strongly confirming that galactic cosmic rays originate in OB associations. As shown in Fig. 14-8, this result also confirms the preferred acceleration of refractory elements as compared to volatile ones and derives a mass dependence of the relative abundances for both refractory ($\propto A^{2/3}$) and volatile ($\propto A$) elements. This result has been confirmed also by measurements at higher energies, up to almost 4 TeV (Ahn et al. 2010).

Spatial diffusion. The diffusion on the irregularities of the galactic magnetic field explains the highly isotropic distribution of CRs and their confinement in the Galaxy. The scatter of charged particles on the random inhomogeneities of the magnetic field B is usually treated in the quasi-linear approximation, according to which the average and fluctuating fields are separated and it is assumed that fluctuations $\delta B \ll B$ are small (B is the regular field). The interaction between the waves and the cosmic particles is of resonant type and is maximized when the irregularities on the magnetic field have the wave vector component parallel to the average magnetic field $K_{\parallel} = \pm s/(r_g \mu)$, where the integer s is the order of cyclotron resonances, $r_g = R/B$ is the gyroradius, $R = pc/Ze$ is the particle rigidity, and μ is the particle pitch angle. According to Ptuskin et al. (1993b),

$$K_{\parallel}(\vec{r}, R) \simeq \frac{1}{3} \nu r_g / P,$$

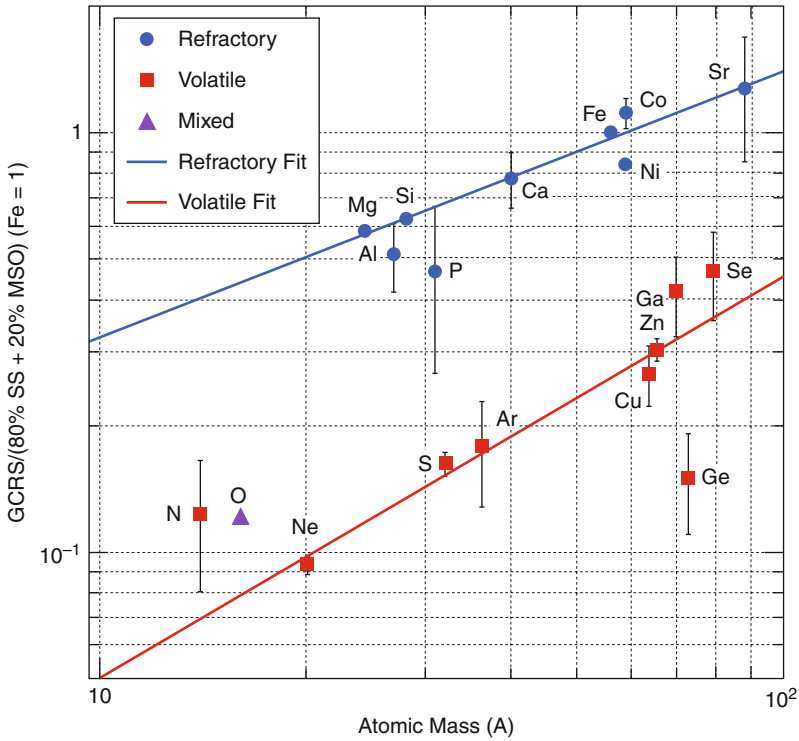
where P is the integral of the normalized power spectrum of the turbulent hydromagnetic fluctuations $P = \int \Delta B(k)^2 / B^2$. It is generally assumed that $B(k)^2 \propto k^{-a}$, with different predictions for a . The Kolmogorov turbulence spectrum is identified by $a = 5/3$, which gives $K_{\parallel} \propto R^{1/3}$.



■ Fig. 14-7

Abundances of refractory nuclides in GCR and solar system (propagation modeled by a leaky-box model) (Wiedenbeck et al. 2003)

An exponent $a = 3/2$ is typical for the Kraichnan turbulent spectrum, corresponding to $K_{\parallel} \propto R^{1/2}$, which is closer to the trend of the diffusion coefficient derived from pure diffusion interpretation of high-energy CR data. The most direct way to deduce phenomenologically the diffusion coefficient is the comparison with data on secondary-to-primary nuclei fluxes. The best data are for the ratio of boron-to-carbon (B/C). The boron is a purely secondary nucleus, produced mostly by the fragmentation of carbon, nitrogen, and oxygen nuclei, while carbon is a primary species directly produced in the SNR. In the case of pure diffusion (neglecting energy losses and gains, convection, and nuclear destructions), $B/C \propto K(E) \propto R^{-\delta}$. The main difference between the two diffusive schemes mentioned above is in that the Kraichnan model predicts a softer (steeper) observed spectrum $\propto R^{-0.5}$ and closer to the observation of the secondary-to-primary ratio (B/C above all) than the Kolmogorov spectrum, which would predict a $B/C \propto R^{-0.3}$ at ≥ 20 – 30 GeV. In the quasi-linear theory, the perpendicular diffusion



■ Fig. 14-8

Ratio of GCR abundances to a mixture of 80% solar system and 20% massive star outflow vs atomic mass A (Link et al. 2009)

coefficient is suppressed with respect to the parallel one, meaning that diffusion takes place mostly along the regular field direction. The quasilinear theory leads to a rigidity power law for the diffusion coefficient, which is usually assumed to have the form:

$$K(E) = K_0 \beta^\eta R^\delta \quad (14.4)$$

(β is the Lorentz factor). K_0 is linked to the level of the hydromagnetic turbulence and δ to the density spectrum of these irregularities at different wavelength. η is usually set to 1, while a different value parameterizes very low-energy deviations. The lack of information on the magnetic field irregularities prevents us from a precise determination of the diffusion coefficient, which is instead possible only from interpretation of cosmic-ray data. The $\delta = 1/3$ case, together with some amount of reacceleration, can reproduce the peculiar boron-to-carbon (B/C) peak observed at $\sim 1 \text{ GeV n}^{-1}$, but it tends to overestimate the data at higher energies, for which $\delta = 0.5\text{--}0.7$ seems preferred. On the other hand, a diffusion coefficient $\delta \sim 0.6$, that is, a short confinement time, would imply a high degree of anisotropy in the very high-energy region above 10^{14} eV , at variance with the results of the experimental measurements (see [Sects. 2.3](#) and [5.1](#)).

Reacceleration. Diffusive, or continuous, reacceleration is a second-order Fermi acceleration process due to the scattering of charged particles on the magnetic turbulence in the interstellar

hydrodynamical plasma. The diffusive reacceleration coefficient is related to the velocity of such disturbances, called the Alfvén velocity V_A , and is naturally connected to the space diffusion coefficient through the relation (Seo and Ptuskin 1994):

$$K_{pp} \times K = \frac{4}{3} V_A^2 \frac{p^2}{\delta(4-\delta^2)(4-\delta)}, \quad (14.5)$$

where p is the particle momentum and δ is the power index of the diffusion coefficient K as in (Eq. 14.4). The scattering centers drift in the Milky Way with a velocity $V_A \sim 20\text{--}100 \text{ km s}^{-1}$. Diffusive reacceleration contributes significantly in shaping the B/C ratio at kinetic energy per nucleon E around 1 GeV n^{-1} . At energies greater than few tens of GeV/n , the energy change due to reacceleration is negligible and the relevant effect can be safely neglected in the propagation equation.

Convection. The stellar activity and the energetic phenomena associated to the late stage of stellar evolution may push the ISM and the associated magnetic field out of the galactic plane. The net effect is likely a convective wind directed outwards from the galactic plane, which adds a convective term to the diffusion equation (Eq. 14.3) (Ipavich 1975; Parker 1965). The properties of this wind may be studied by hydrodynamical (Johnson and Axford 1971; Zank 1989) and magneto hydrodynamical (MHD) (Schlickeiser and Lerche 1985) methods. Convection may dominate at low energies, depending on its intensity V_C and if δ is large, and compete with diffusion up to few tens of GeV/n , but its role becomes negligible at higher energies.

Energy losses. Nuclei lose energy via ionization in the ISM neutral matter (90% H and 10% He) and Coulomb interaction in a completely ionized plasma, dominated by scattering off the thermal electrons. The expressions for the corresponding energy loss rates can be found in Mannheim and Schlickeiser (1994) and Strong and Moskalenko (1998).

Nuclear fragmentation. Charged CR nuclei heavier than protons suffer catastrophic losses due to nuclear destruction on the interstellar H and He. These reactions become irrelevant on the CR spectrum at energies $\gtrsim 1 \text{ TeV n}^{-1}$, depending on the nuclear species. Inelastic interactions lower and flatten the flux of the broken nucleus, while giving rise to the secondary nuclei produced in the inelastic collision. Fragmentation cross sections can be derived from the semiempirical formulation of Webber et al. (1990) (see also Maurin et al. 2001 and references therein).

2.2 Experimental Methods for Direct Measurements

Direct measurements of the galactic cosmic rays are possible up to $\simeq 10^{14} \text{ eV}$ by means of experiments flown on balloons or satellites.

- *Balloon-borne experiments* can be assembled with a moderate budget, and in general allow multiple measurements, in that their payload can be eventually recovered, repaired, and flown again. On the other hand, they can provide only a limited exposure of few days of flight, due to the limited resources onboard and to the winds, which direction and velocity can drive balloons far from the launch site or above populated areas.

This limit is constantly being pushed forward with the recent advent of the so-called long duration balloon flights, where balloons with volume $\gtrsim 10^6 \text{ m}^3$ and suspended weight $\simeq 3\text{--}4 \text{ t}$ can fly for tens of days. The CREAM experiment (Seo et al. 2004) reached the record flight duration of 42 days in 2005 in Antarctica. An intense activity of research and development is

ongoing to produce ultra-long duration balloons and pumpkin-shaped superpressure balloons using a very thin closed plastic skin, able to stay aloft up to 100 days. Prototypes were successfully flown in Antarctica in the last couple of years.

Two different classes of experiments on balloons have been realized so far, the first searching for antimatter in cosmic rays (Barwick et al. 1997a; Boezio et al. 2001; Wang et al. 2002; Yoshida et al. 2004) and the second aiming at the measure of the cosmic ray primary composition up to 100 TeV (Chang et al. 2008; Müller et al. 2009; Seo et al. 2004).

- *Satellite experiments* have longer exposure and can avoid the background related to the residual atmosphere above the balloons, but their cost is very high. They can fly at different orbits: polar ones, to study low-energy cosmic rays at high latitudes, or equatorial ones if they want to detect gamma radiation, to be screened from cosmic rays. Research activity on charged cosmic rays is performed both to search for antimatter and to study chemical abundances (Adriani et al. 2009a). Typical detectors employ various combinations of different instruments, which aim is that of measuring the magnitude of the incoming particle charge and its energy. The goal of measuring the single primary elements, both in spectrum and abundance, requires adequate exposures: for example, instruments aiming at reaching the knee energies with minimal statistics of about 10 events above 10^{15} eV need exposures from ≈ 1 to $20 \text{ m}^2 \text{ sr y}$ for H and heavy elements, respectively.
- *Space-based experiments* can be hosted onboard space stations, with stricter requirements and higher costs. The AMS detector has been designed to operate as an external module on the International Space Station, with the aim of searching for antimatter and dark matter while performing precision measurements of cosmic-rays composition and flux (Zuccon et al. 2009). The Jem-EUSO detector is planned to be located on the Japanese module of the space station to study the extreme high-energy phenomena in the universe, at $E > 10^{20}$ eV (Takahashi et al. 2009).

2.2.1 The Measure of Energy

In magnetic spectrometers, the rigidity $R = pc/Ze$ and the sign of the charge of the crossing particles can be determined. Measuring the rigidity R as a function of the radius of curvature of the particle, $R = B r_{\text{curv}}$ in a magnetic field B , and the charge from the ionization energy lost in the tracker, the particle momentum can be obtained. The performance of the spectrometer is characterized by the distribution of the maximum detectable rigidities MDR for all trajectories, which is generally defined such that the error in the measure of R be

$$\frac{\delta R}{R} = \frac{R}{\text{MDR}}.$$

The sign of the charge of a particle can be reliably determined for rigidities up to $1/3$ MDR.

Calorimeters or transition radiation detectors are used to derive the particle energy. The first are generally employed for the measure of low-charge particles such as protons and helium nuclei. Due to the limitations in thickness and weight inherent to their location on balloons or satellites, they cannot be used to study higher Z nuclei. If E_1 is the energy required to produce a detectable “quantum” of energy in the detector (a photon for a scintillator, a positron-electron pair for a semiconductor device, an electron-ion pair for a gas array, etc.) and E_0 is the initial energy, then $N = \epsilon E_0/E_1$ is the actual number of recorded quanta (ϵ being the collection and detection efficiency of the sensitive element). The energy resolution is dominated by Poissonian

fluctuations and improves with increasing primary energy, as well as in detectors with smaller E_1 (e.g., $E_1 \simeq 1$ eV for semiconductors, while $E_1 \simeq 100$ eV for scintillators). Typically, thick targets with sufficient interaction lengths to force the interaction of hadrons are placed in front of an electromagnetic calorimeter, thick enough in radiation lengths to fully absorb the secondary cascade. Calorimeters can be calibrated at accelerator beams with protons and heavy ions.

For nuclei with $Z > 3$, transition radiation detectors (TRD) can measure the Lorentz factor $\gamma_L = E/m$, which together with the knowledge of the particle mass can provide an energy measurement. This radiation is emitted in the X-ray region when a particle traverses the boundary between two media with different dielectric properties and is proportional to γ_L . It is observable above $\gamma_L \gtrsim 400$, where all other methods already give saturated signals, up to $\gamma_L \simeq 50,000$. The upper limit is a consequence of the interference in emission from the multiple layers of material which constitute the radiator; in fact, multiple foils are needed to enhance the probability of emission, which for a single charge is of the order of $1/137$.

The TRD has the advantage of having a relatively low mass; thus, larger collecting areas can be set on balloons or satellites despite the weight limitations. On the other hand, no measurement for $Z < 3$ is possible, due to the low photon number and consequently the too large fluctuations. When both TRD and calorimeters are used, a fraction of the incoming particles can cross both of them, thus providing a cross calibration of the energy determination, at least for $Z > 2$, with completely different systematic uncertainties. Cherenkov counters and proportional tubes are also used, offering together with the TRD a response proportional to the Z^2 ; differently from magnetic spectrometers, the misidentification of the charge can heavily influence the determination of the Lorentz factor and hence of the energy.

At higher energies, emulsion chambers are used to study the angle and energy of the electromagnetic cascades from the decays of the neutral pions produced in the first interaction of the primary in the detector (Asakimori et al. 1998; Derbina et al. 2005). Good accuracies are obtained in the measurement of the energy released in the electromagnetic component in such interaction (about 15%); the largest source of uncertainty is represented by the fluctuations in the fraction of primary energy going into π^0 , which decreases for heavier primary nuclei.

2.2.2 The Measure of the Charge

The measure of the charge is generally based on the use of scintillators or solid-state detectors. The ionization energy lost by the particle is proportional to Z^2 . Combining the measure of the kinetic energy and of the energy loss in a thin detector, the atomic number of the particle can be obtained. For antimatter studies, the rigidity, charge, and charge sign of a particle can be measured by means of magnetic spectrometers, consisting of a magnet (either permanent or superconducting) and a tracker. In Cherenkov counters, the signal of a particle with charge Z and velocity β (in units of c) is $\propto Z^2(1 - 1/\beta^2 n^2)$, with n = refractive index of the medium; its response abruptly drops below a threshold Lorentz factor given by $\beta = 1/n$.

Different detectors can be employed together to measure the charge in a redundant way. For example, four independent instruments are used in the CREAM detector (Seo et al. 2004): a timing-based scintillator charge detector, a plastic Cherenkov detector, a scintillating fiber hodoscope, and a silicon charge detector. The use of detector segmentation or timing technique allows backslash particles produced, for example, in a calorimeter in the lower part of the payload to be tagged.

The required resolution in charge must reach 0.2 charge units or better to resolve different elements (e.g., boron from carbon, for which the flux ratio can be as low as 1%).

At energies above 1 TeV, emulsion chambers are also used, where the particle charge is measured through the ionization produced in stacks of nuclear or X-ray emulsion plates; the “darkness” of the spots produced when a charged particle crosses the sensitive layers is proportional to the nuclear charge.

2.3 Solutions to the Diffusive Equation and Comparison with Data

The solution of the master diffusion equation (Eq. 14.3) has been deeply investigated, and several different techniques proposed in the literature can lead to similar fluxes at the Earth, at least for stable nuclei (Jones et al. 2001; Maurin et al. 2002b; Strong et al. 2007).

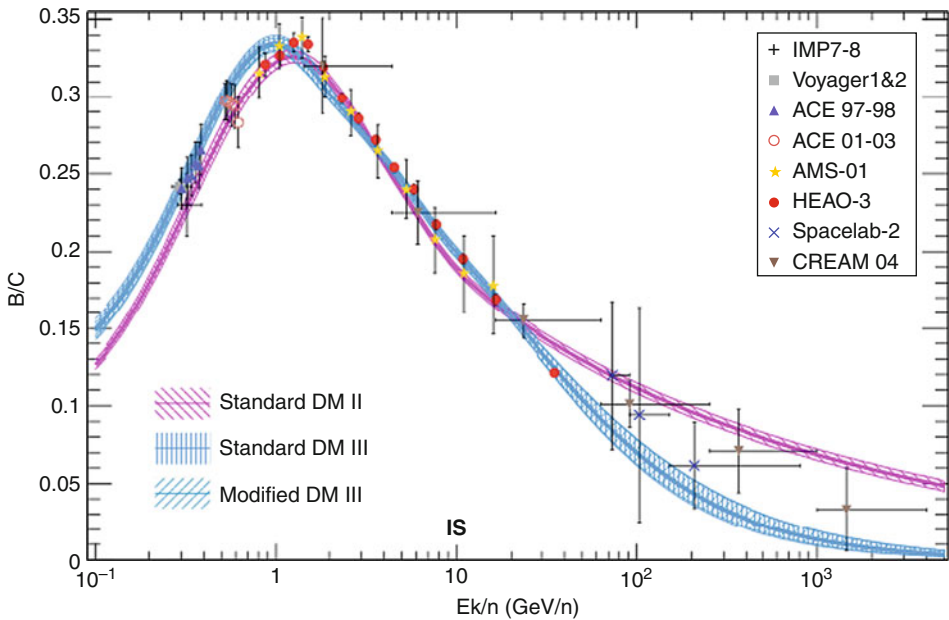
In diffusive models (Ginzburg et al. 1980; Webber et al. 1992), cosmic rays diffuse in a region called diffusive halo (DH), pictured as a thick disk high somehow between 1 and 15 kpc, which matches the circular structure of the Milky Way. The galactic disk of stars and gas, where primary cosmic rays are accelerated, lies in the middle. It extends radially 20 kpc from the center and has a half-thickness h of 100 pc.

Fully numerical solution is the one adopted by the GALPROP code (Strong and Moskalenko 1998; Trotta et al. 2011), which is particularly useful if aiming at the production of gamma rays from charged cosmic rays. Gamma rays do not diffuse and are therefore observed at their source, so that a full spatial treatment is required (Strong et al. 2004). Comparable results for the propagation of stable primary and secondary nuclei have been obtained with the similar Dragon code (Evoli et al. 2008) and with the modified weighted slab technique (Jones et al. 2001).

The Bessel expansion method, based on the cylindrical symmetry of the DH and on approximate values for the ISM (not relevant for charged CR propagation), allows a two-dimensional fully analytical model (Donato et al. 2002; Maurin et al. 2001, 2002a, b). Numerical solution is required only for the diffusion in energy space. If the interest is concentrated on charged particles, the complete spatial solution is often redundant, since they diffuse for long time before being detected and they do not backtrace their source.

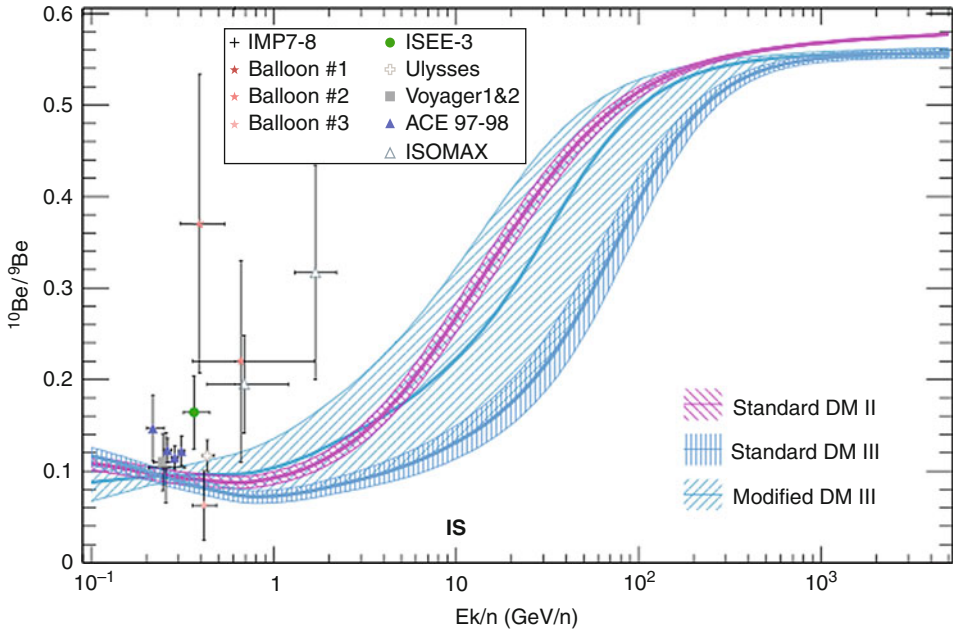
The free parameters of diffusive models – typically the normalization and power spectrum index of the diffusive coefficient K_0 and δ and the thickness of the DH, convective, and Alfvén velocities V_C and V_A – are fixed by studying the B/C ratio which is quite sensitive to cosmic-ray transport and which may be used efficiently as a constraint. These parameters were self-consistently constrained in Maurin et al. (2001) and Putze et al. (2010). When only B/C is considered, models with diffusion, convection, and reacceleration are preferred over the diffusion and reacceleration case. The former points towards $\delta \sim 0.7$ – 0.8 , while the latter points towards $\delta \sim 0.3$. This result does not depend on the halo-size L . Moreover, a B/C analysis based on AMS-01 data (instead of HEAO-3) indicates that the presence of convection and reacceleration is required and points to a diffusion slope $\delta \sim 0.5$, closer to theoretical expectations. As regards the normalization of the diffusion coefficient K_0 , it is worth noting that B/C measurements actually constrain K_0/L , not K_0 alone (Maurin et al. 2001; Trotta et al. 2011). However, no definite answer is given to the precise values of the propagation parameters – δ being the most relevant one – and a high degeneracy among the parameters is observed. Data from PAMELA and from AMS02 are awaited to help solving the long-standing issue on the value for δ .

One possible improvement to the understanding of the propagation physics would be the simultaneous study of stable and radioactive species – usually believed to be powerful tracers of the diffusive halo size. In fact, the combined analysis does not yet allow to fully break this degeneracy (Donato et al. 2002). Secondary radioactive species, such as ^{10}Be , originate from the spallation of their heavier progenitors with the ISM and have half-lives τ_0 ranging from 0.307 My for ^{36}Cl to 1.51 My for ^{10}Be . The quantity $l_{\text{rad}} = \sqrt{K\gamma\tau_0}$ is the typical distance on which a radioactive nucleus diffuses before decaying. Using $K \approx 10^{28} \text{ cm}^2 \text{ s}^{-1}$ and $\tau \approx 1 \text{ My}$, the diffusion length is $l_{\text{rad}} \approx 200 \text{ pc}$. Therefore, any underdense region on a scale $r_h \sim 100 \text{ pc}$ around the Sun (as the observed Local Bubble) leads to an exponential attenuation of the flux of radioactive nuclei. The attenuation is energy and species dependent (Donato et al. 2002). In diffusion/convection/reacceleration models, $L \sim 8 \text{ kpc}$ with $r_h \sim 120 \text{ pc}$ (Putze et al. 2010), and in diffusion/reacceleration models, $L \sim 4 \text{ kpc}$ and no underdense region is necessary to explain the data (Putze et al. 2010; Trotta et al. 2011). The halo size comes out as an increasing function of the diffusion slope δ . A striking feature is that in many models, r_h points to $\sim 100 \text{ pc}$, a value supported by direct astronomical observations of the local ISM (see discussions in Donato et al. 2002; Putze et al. 2010 and references therein). In \blacktriangleright Figs. 14-9 and \blacktriangleright 14-10, we display results obtained on a fit to B/C and $^{10}\text{Be}/^9\text{Be}$ data. The free parameters of the model, namely, $L, K_0, \delta, V_C,$ and V_A have been varied in very large ranges. The results demonstrate the above considerations on the degeneracy among the free parameters of the diffusive models. A good fit is possible up to tens of GeV/n, where the data are good as well. The figure relevant to the



\blacksquare Fig. 14-9

68%CL envelopes (*shaded areas*) and best fit (*thick lines*) for diffusive/reacceleration models ($r_h = 0$, *red*) and for diffusive/reacceleration/convection models (*blue*) tested on B/C data (From Putze et al. (2010))



■ Fig. 14-10

68% CL envelopes (*shaded areas*) and best fit (*thick lines*) for diffusive/reacceleration models ($r_h = 0$, *red*) and for diffusive/reacceleration/convection models (*blue*) on $^{10}\text{Be}/^9\text{Be}$ (From Putze et al. (2010))

$^{10}\text{Be}/^9\text{Be}$ ratio also indicates the paucity of data for radioactive isotopes (the same is even more evident for all other radioactive isotopes).

A complementary view to cosmic-ray propagation is given by the anisotropy considerations. A high isotropy is peculiar for galactic cosmic rays. Indeed, the global leakage of particles from the Galaxy and the contribution of local sources can lead to an anisotropy, but the regular and stochastic magnetic fields tend to isotropize the angular distribution of cosmic rays. Only at high energies it is possible to detect some deviation from pure isotropy (see also discussion in Sect. 5.1). The amplitude of the anisotropy, namely, the gradient of the density of cosmic rays, δ_{AN} , may be written as $\delta_{\text{AN}} = -[3K(E)]\nabla N/\nu N$ in the case of pure isotropic diffusion, where $K(E)$ is the diffusion tensor (Berezinskii et al. 1990) and ν is the velocity of the cosmic ray. The calculations indicate that a diffusion coefficient $K(E) \propto R^{0.3}$ (the Kolmogorov spectrum) is compatible with data on cosmic-ray anisotropy within a factor of about 3, while a stronger dependence on energy as $K(E) \propto R^{0.6}$ (which is closer to observed absolute nuclei fluxes) tends to overpredict anisotropy at $E \geq 10^{14}$ eV (Blasi and Amato 2011).

2.4 Antimatter in CRs

The presence of a small amount of antimatter in cosmic rays is predicted from spallation reactions of incoming protons and helium nuclei on the ISM (contributions from cosmic rays

with higher Z being negligible). The spallation products of these inelastic scatterings account for quarks and gluons, which immediately hadronize. In particular, the hadronization process includes the production of a small amount of antiprotons, antideuterons, and positrons (the latter induced by decay processes). Antiprotons and positrons in cosmic rays have been measured in recent years with increasing accuracy and in an energy window spanning from few hundreds of MeV to hundreds of GeV. For antideuterons, only upper limits have been set.

The search for cosmic antimatter is a further test of the propagation model. The study of light antimatter, due to its tiny flux, is optimal to search for contributions with spectral peculiarities, such as nonthermal production or annihilation from dark matter (DM) particles in the galactic dark halo. The latter are usually predicted to be weakly interacting massive particles (WIMPs).

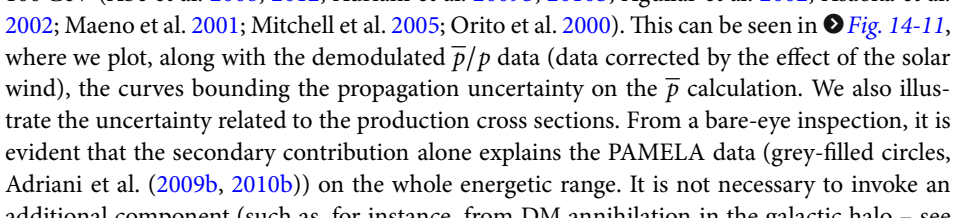
The indirect dark matter detection is based on the search for anomalous components due to the annihilation of DM pairs in the galactic halo, in addition to the standard astrophysical production (Bertone et al. 2005; Bottino et al. 2004; Salati et al. 2010). These contributions are potentially detectable as spectral distortions in various cosmic antimatter fluxes, as well as γ -rays and ν 's:

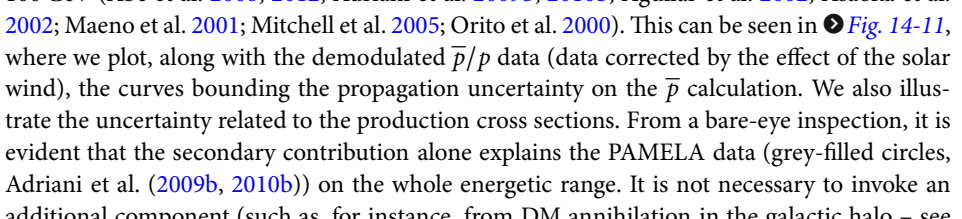
$$\chi + \chi \rightarrow q\bar{q}, W^+W^-, \dots \rightarrow \bar{p}, \bar{D}, e^+ \gamma \& \nu's. \quad (14.6)$$

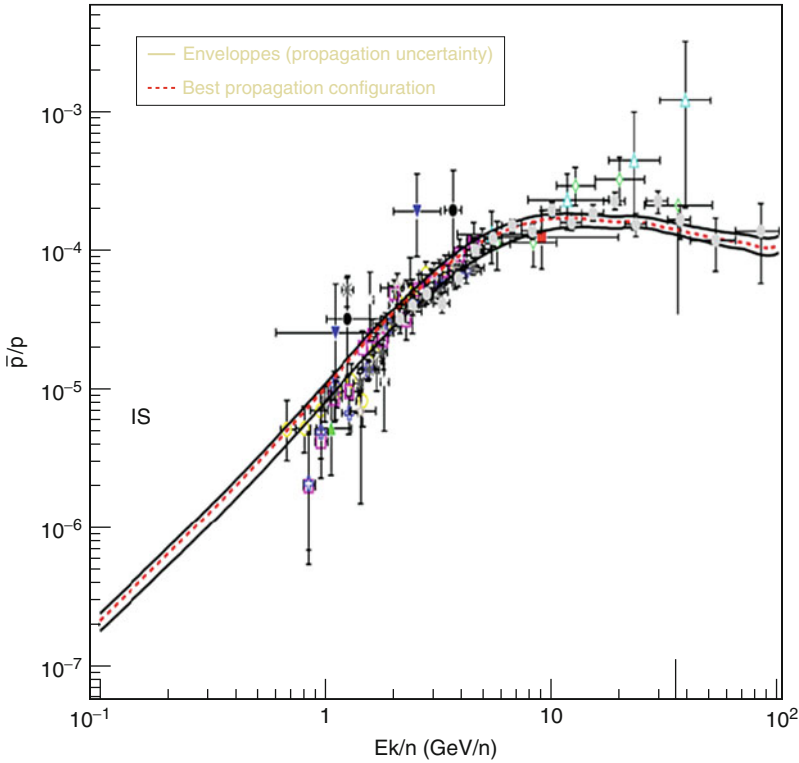
The search for exotic signals in cosmic antimatter is motivated by the very low-astrophysical counterpart, acting as a background. Detection of the DM annihilation products has motivated the spectacular development of several new experimental studies.

2.4.1 Antiprotons

Antiprotons may be produced by spallation of high-energy primary nuclei impinging on the atoms of the ISM inside the galactic disk (Bergstrom et al. 1999; Donato et al. 2001). They represent the background when searching for small peculiar contributions such as signals from DM annihilation.

The secondary antiproton flux has been predicted with small theoretical uncertainties (Donato et al. 2001, 2009) and reproduces astonishingly well the data from 200 MeV up to at 100 GeV (Abe et al. 2008, 2012; Adriani et al. 2009b, 2010b; Aguilar et al. 2002; Asaoka et al. 2002; Maeno et al. 2001; Mitchell et al. 2005; Orito et al. 2000). This can be seen in  [Fig. 14-11](#), where we plot, along with the demodulated \bar{p}/p data (data corrected by the effect of the solar wind), the curves bounding the propagation uncertainty on the \bar{p} calculation. We also illustrate the uncertainty related to the production cross sections. From a bare-eye inspection, it is evident that the secondary contribution alone explains the PAMELA data (grey-filled circles, Adriani et al. (2009b, 2010b)) on the whole energetic range. It is not necessary to invoke an additional component (such as, for instance, from DM annihilation in the galactic halo – see below) to the standard astrophysical one.

The annihilation of DM candidate particles throughout the whole Milky Way volume may generate primary antiprotons. In this case, the WIMP annihilations take place all over the diffusive halo. The antiproton signal from annihilating DM particles leads to a primary component directly produced throughout the DH (Donato et al. 2004). The variation of the astrophysical parameters induces a much larger theoretical uncertainty on the primary than on the secondary flux: in the first case, the uncertainty reaches two orders of magnitude for energies $T_{\bar{p}} \lesssim 1$ GeV, while in the second case, it never exceeds 25% (see  [Fig. 14-11](#) for the secondary population;



■ Fig. 14-11

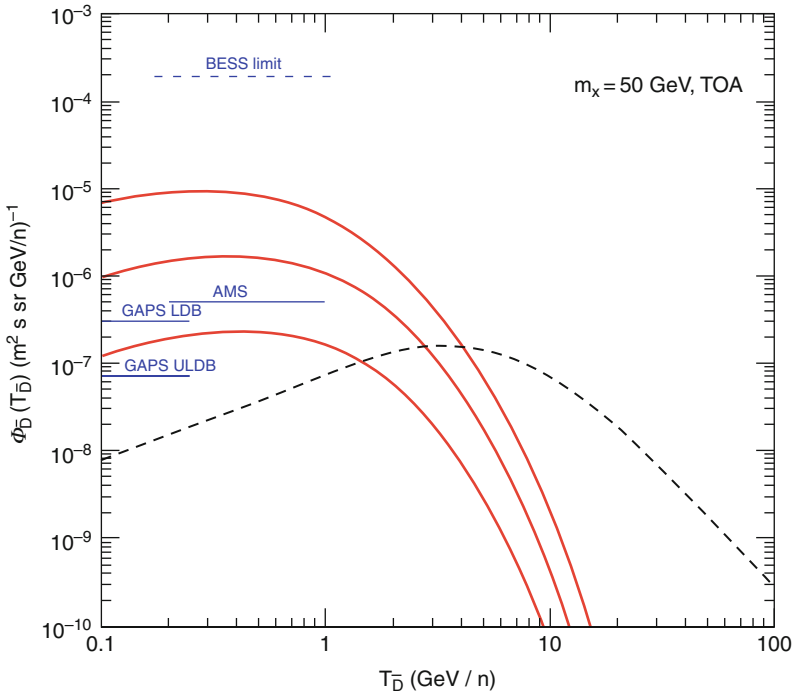
Propagation uncertainty envelopes of the IS \bar{p}/p ratio and two parameterizations of the production cross section (Donato et al. 2009)

for the primary DM component, see the discussion on [Fig. 14-12](#)). The reason is in the location of the sources: the primary flux due to dark matter annihilation originates in the whole diffusive halo and is very sensitive to the halo size (which is varied between 1 and 15 kpc) and the convective velocity. Therefore, it is of the utmost importance to constrain the propagation parameters in order to evaluate any possible exotic contribution to the \bar{p} flux from sources in the whole halo.

2.4.2 Antideuterons

It was shown that the antideuteron spectra deriving from DM annihilation are expected to be much flatter than the secondary astrophysical component at low kinetic energies, $T_{\bar{d}} \lesssim 2 - 3 \text{ GeV n}^{-1}$, thus offering a potentially very clear indirect detection channel (Donato et al. 2000, 2008). Antideuterons have not been measured so far, and only an upper limit has been obtained in (Fuke et al. 2005).

The secondary \bar{d} flux is the sum of the 6 contributions corresponding to p , He, and \bar{p} cosmic-ray fluxes impinging on H and He IS gas (other reactions are negligible) (Chardonnet et al. 1997;



■ Fig. 14-12

Antideuteron flux for dark matter sources (*solid*) and secondary contribution (*dashed*). *Horizontal solid lines* show the estimated sensitivities for next-generation experiments (Donato et al. 2008)

Duperray et al. 2005). The production cross sections for these specific processes are those given in Duperray et al. (2005). The solution to the propagation equation has the same expression as for secondary antiprotons (Donato et al. 2008).

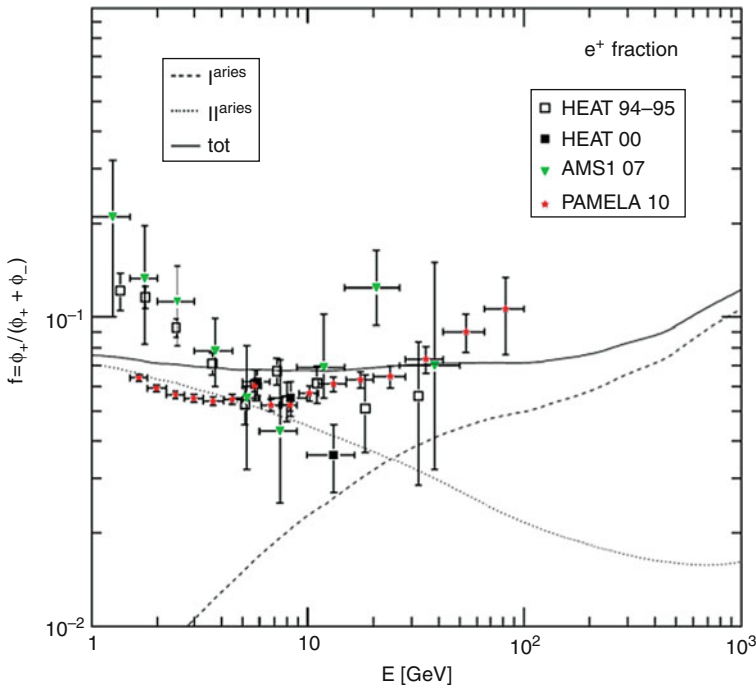
The production of cosmic antideuterons is based on the fusion process of a \bar{p} and \bar{n} pair – being the pair produced from CRs collisions or from DM annihilation in the dark galactic halo. One of the simplest but powerful treatment of the fusion of two or more nucleons is based on the so-called coalescence model which, despite its simplicity, is able to reproduce remarkably well the available data on light nuclei and antinuclei production in different kinds of collisions (Donato et al. 2000, 2008; Duperray et al. 2005 and references therein).

We present in ► Fig. 14-12 a possible experimental scenario. The secondary \bar{d} flux is plotted alongside the primary flux from DM particle ($m_\chi = 50 \text{ GeV}$) annihilating in the halo. The three curves bound the propagation uncertainties which – as for the case of \bar{p} from galactic DM, grossly rescaled by a factor 10^3 – span almost two orders of magnitude on the whole energetic range. The present BESS upper limit (Fuke et al. 2005) is at a level of $2 \cdot 10^{-4} (\text{m}^2 \text{ s sr GeV n}^{-1})^{-1}$. We also plot the estimated sensitivities of the gaseous antiparticle spectrometer GAPS on a long duration balloon flight (LDB) and an ultralong duration balloon mission (ULDB) (Aramaki et al. 2010) and of AMS-02 for 3 years of data taking. The perspectives to explore a part of the parameter space where DM annihilations are mostly expected (i.e., the low-energy tail) are promising (Donato et al. 2008).

2.4.3 Positrons

Secondary positrons are produced – like antiprotons and antideuterons – by the spallation of the interstellar medium when impinging high-energy particles (Delahaye et al. 2009; Moskalenko and Strong 1998). The main production channel is the collision of protons with hydrogen atoms at rest producing charged pions π^\pm which decay into muons μ^\pm . The latter are also unstable and eventually lead to electrons and positrons. Positrons may also be produced through kaons although this channel is rare. In the case of positrons and electrons (► Eq. 14.3), describing the propagation of cosmic rays throughout the DH is dominated by space diffusion and energy losses. Above a few GeV, synchrotron radiation in the galactic magnetic fields as well as inverse Compton scattering on stellar light and on CMB photons dominates.

Data on the absolute e^+ flux are less precise than on \bar{p} (Adriani et al. 2010a, 2011; Alcaraz et al. 2000; Barwick et al. 1997b; Boezio et al. 2000). Nevertheless, the data are well described by the contribution from spallation reactions within experimental error bars. The positron fraction $e^+/(e^+ + e^-)$ has been measured by the PAMELA satellite experiment (Adriani et al. 2009c, 2010a) and confirmed by the Fermi-LAT collaboration discriminating the lepton charge using the Earth magnetic field (Ackermann et al. 2012). It increases with energies, at variance with the predictions from pure secondary production of cosmic positrons. A viable explanation of the experimental result resides in the additional contribution of astrophysical sources accelerating leptons in their sites (Blasi and Serpico 2009; Delahaye et al. 2010). This result is illustrated in ► Fig. 14-13, where the $e^+/(e^+ + e^-)$ has been



■ Fig. 14-13

Template calculation for the positron fraction including all primary (discrete local and smooth distant) and secondary electrons and positrons (Delahaye et al. 2010)

calculated adding to the secondary production (mostly relevant for positrons) the contributions from standard astrophysical sources, such as supernova remnants and pulsars. As shown in Delahaye et al. (2010) (and references therein), the cosmic fluxes of positrons and leptons are quite sensitive to the presence of sources in the near Galaxy (few kpc), whose physics could be explored also in this peculiar channel. The most recent experimental data on $e^+ + e^-$ spectrum are displayed in [Fig. 14-14](#) (Ackermann et al. 2010). The $e^+ + e^-$ spectrum has been computed with a GALPROP model (shown by solid black line) with breaks in the acceleration power spectrum and a strong cutoff above 2 TeV. Blue lines show e^- spectrum only. The dashed/solid lines show the before modulation/modulated spectra. Secondary e^+ (red lines) and e^- (orange lines) are also shown. Secondary electrons and positrons from CR proton and helium interactions with interstellar gas make a significant contribution to the total lepton flux, especially at low energies. The total leptonic flux is characterized by peculiar spectral features which could be explained by local astrophysical or exotic sources in a few kpc region around the solar system (see Ackermann et al. 2010; Delahaye et al. 2010 and references therein for details).

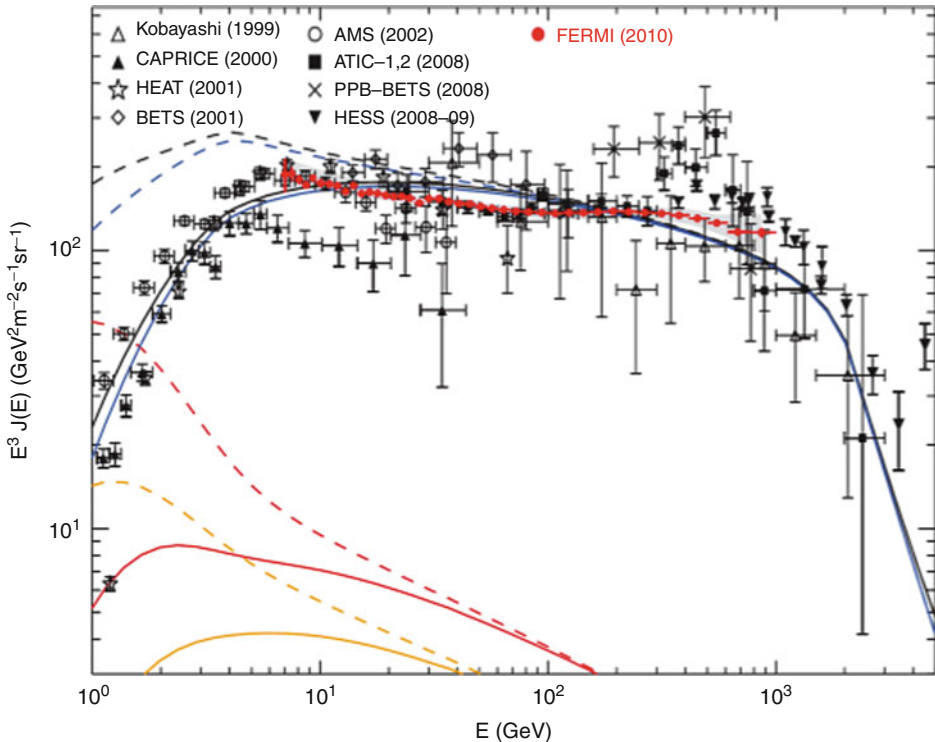


Fig. 14-14

The measured $e^+ + e^-$ spectrum along with a possible estimation of primary and secondary leptonic components (figure taken from Ackermann et al. (2010))

3 From 100 TeV n⁻¹ to 100 PeV n⁻¹

3.1 Extensive Air Showers

Primary cosmic rays above $\approx 10^{14}$ eV are characterized, as shown in [Fig. 14-1](#), by a low flux, and their energy, mass, and arrival directions can be studied only indirectly by exploiting the particle cascades that they produce in the atmosphere. The measured observables are the longitudinal and lateral distributions of the charged components or the Cherenkov and the fluorescence light produced during the propagation of the extensive air shower (EAS) in the Earth atmosphere down to the experimental level. Since all observables are interrelated and depend in different ways on both energy and mass of the primaries, multiparametric measurements are to be preferred: modern experimental setups in fact include detectors of many shower components.

When a primary cosmic-ray nucleus interacts with an air nucleus in the upper atmosphere, a leading nucleon emerges, while a fraction (the so-called inelasticity k) of its initial energy goes into production of secondaries,² mainly π mesons; due to charge independence, the energy is equally shared among π^+ , π^- , and π^0 .

The electromagnetic component (electrons³ and photons) originates from the fast decay of neutral pions into photons, which initiates a rapid multiplication of particles in the shower, mainly through two production processes: bremsstrahlung by electrons and pair production of electrons by photons. The multiplication continues until the rate of energy loss by bremsstrahlung equals that of ionization, at a critical energy which in air is $E_c \approx 86$ MeV. The hadronic back-bone of the shower continuously feeds the electromagnetic part; the charged pions can either interact or decay. The nucleon interaction length in air (with $\langle A \rangle \approx 14.5$) is ≈ 80 g cm⁻². The transverse momentum of nucleons and pions and the multiple scattering of the shower particles, particularly of the electrons, are responsible for the lateral spread of the particles in the shower. Finally, charged pions decay into muons (and neutrinos). Since muons lose energy mainly through ionization and excitation, they are not attenuated very much and give rise to the most penetrating component of the shower.⁴ A sketch of the different components of an extensive air shower is shown in [Fig. 14-15](#) (left).

Transport and cascade equations describe the development and propagation of the EAS in the atmosphere. Path lengths are generally measured in units of g cm⁻² to remove the effect of the density of the medium: the vertical atmospheric depth is

$$X_V(h) = \int_h^\infty \rho(h') dh' \text{ g cm}^{-2}, \quad (14.7)$$

where $\rho(h)$ is the density of the atmosphere at altitude h . The atmospheric depth measured downward from the top along the direction of the incident particle is the *slant depth* X . They are sketched in [Fig. 14-15](#) (right).

²The definition of “secondaries” applies here to cosmic rays produced in the Earth atmosphere, not to be confused with the secondary particles originating from the spallation of primary cosmic rays in the interstellar medium.

³From now on, “electrons” stay for both e^- and e^+ .

⁴This is why, despite being electromagnetic particles, muons are traditionally not included in the electromagnetic shower component, but in the separated muonic one.

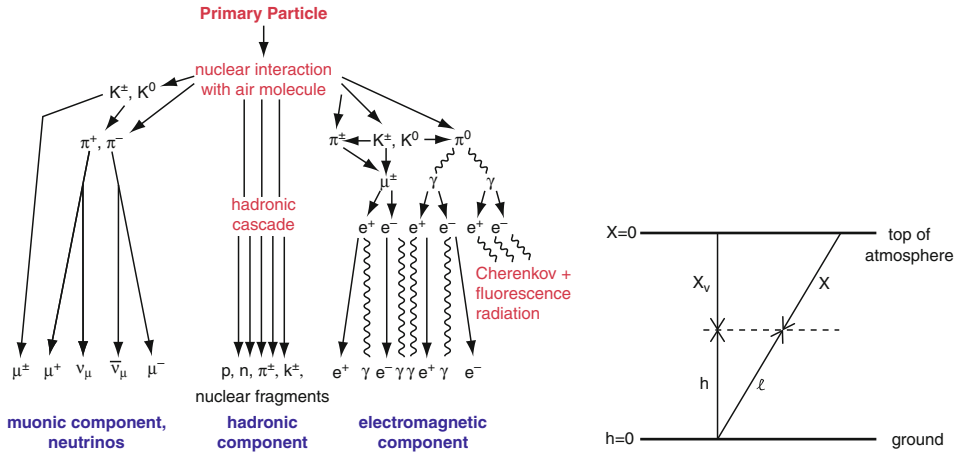


Fig. 14-15

Left: Sketch of an extensive air shower (Haungs et al. 2003). Right: The variables used to describe the atmospheric depths (Gaisser 1990)

3.1.1 The Electromagnetic Component

The number of charged particles (the shower size N_e) at each atmospheric depth X can be derived from the solutions of the cascade equations (Rossi and Greisen 1941). Most recently, the shower profiles are characterized in data analysis using (Gaisser and Hillas 1977)

$$N_e(X) = N_e^{\max} \left(\frac{X - X_1}{X_{\max} - X_1} \right)^{(X_{\max} - X_1)/\lambda} \exp \left(\frac{X_{\max} - X}{\lambda} \right), \quad (14.8)$$

where X and X_1 are the depths of observation and of first interaction and λ is the attenuation length. The depth of maximum development of the shower, X_{\max} is obtained by determining the number of lengths needed for each particle to reach the critical energy E_c : $X_{\max} \simeq X_R \left[\ln \left(\frac{E_0}{E_c} \right) + \alpha \right]$, where E_0 is the primary energy, the radiation length X_R depends on the composition of the medium ($X_R \simeq 37 \text{ g cm}^{-2}$ in air), and α depends on the primary (electron or photon). The rate of variation of X_{\max} per decade of primary energy, $(dX_{\max}/d\log_{10} E)$, is called the *elongation rate*.

The density of the N_{ch} charged particles as a function of the distance from the shower core (the intersection of the shower axis with the ground) is described by the Nishimura-Kamata-Greisen function (Greisen 1960; Kamata and Nishimura 1958)

$$\rho_{\text{ch}}(r) = \frac{N_{\text{ch}}}{2\pi r_M^2} C(s) \left(\frac{r}{r_M} \right)^{s-2} \left(1 + \frac{r}{r_M} \right)^{s-4.5} \quad (14.9)$$

and depends on the multiple Coulomb scattering of electrons. Here, s is the shower *age*, which describes the shape of the distribution ($s = 1$ at shower maximum); $C(s) = \Gamma(4.5 - s)/\Gamma(s)\Gamma(4.5 - 2s)$, with Γ = gamma function.

The root-mean-square scattering (the width of the approximate Gaussian projected angle distribution) undergone by an electron of initial momentum p (MeV/c) as it traverses a thickness x of material is $\theta = (21\text{MeV}/p\beta)\sqrt{x/X_0}$; at the critical energy, the scattering angle for an

electron crossing one radiation length X_0 is $\theta \simeq 14^\circ$. One Moliere unit, $r_M = 21\text{MeV}/E_c$, is the lateral distance by which an electron with $E = E_c$ is scattered as it traverses $1 X_0$ ($r_M \simeq 0.25X_0$ in air). It is the distance encompassing 90% of the shower energy.

At the highest energies, the electromagnetic shower development undergoes modifications due to two competing effects: (a) above 10^{18} eV, the Landau-Pomeranchuk-Migdal effect (Stanev et al. 1982), for which particle production is suppressed in certain kinematic regions, leading to an increase of the shower to shower fluctuations (due to the stochastic development of the cascades) and to deeper maximum; (b) above $10^{19.5}$ eV, the interaction of gamma rays with the geomagnetic field of the Earth (Eerber 1966): in this case, magnetic bremsstrahlung causes the shower to behave as a superposition of hundreds of lower energy showers and the shower to shower fluctuations are significantly reduced.

3.1.2 The Hadronic and Muonic Components

In a hadronic shower, the first interaction happens at $X_0 = \lambda_I \ln 2$, where λ_I is the interaction length of strongly interacting particles, and about 1/3 of the produced charged particles are π^0 . The attenuation length for the hadronic component is larger than λ_I , due to secondary particle production, and is on average $\simeq 120 \text{ g cm}^{-2}$.

The depth of maximum development of the shower after X_0 is thus the same as for an electromagnetic shower of energy $E_0/3N_{\text{ch}}$, so that (Matthews 2005)

$$X_{\text{max}}^p = X_0 + X_R \ln[E_0/3N_{\text{ch}}E_c^\pi] = X_{\text{max}}^{\text{EM}} + X_0 - X_R \ln(3N_{\text{ch}}). \quad (14.10)$$

As a consequence, hadronic showers cannot have higher elongation rate than electromagnetic ones. The reduction is due to the increase of both multiplicity of charged particles and cross section for the hadronic showers. For protons, the elongation rate is $\simeq 58 \text{ g cm}^{-2}$ per decade of energy, as estimated from calculations that model the shower development using the best estimates of the relevant features of the hadronic interactions. Muons are produced in each of the n generations of a hadronic shower when any of the N_{ch} charged particles have energy equal than some decay energy E_c^π . The total number of muons is

$$N_\mu \propto (N_{\text{ch}})^n = \left(\frac{E_0}{E_c^\pi}\right)^\beta, \quad (14.11)$$

where $\beta \simeq 0.85/0.92$ depending on the hadronic interaction model used. The critical pion energy $E_c^\pi \simeq 20 \text{ GeV}$ in a shower generated by a 1 PeV proton.

Due to energy conservation, $E_0 = E_{\text{EM}} + E_{\text{had}}$, where $E_{\text{had}} = N_\mu E_c^\pi$. The fraction of primary energy going into the electromagnetic component is $E_{\text{EM}}/E_0 = 1 - \left(\frac{E_0}{E_c^\pi}\right)^{\beta-1}$, which can be approximated to a power law $E_{\text{EM}}/E_0 \simeq a \frac{E_0}{E_c^\pi}^b$.

Since the electromagnetic size can be expressed as $N_{\text{EM}} \propto E_0/E_c^\pi$, by series expansion around $\frac{E_0}{E_c^\pi} \simeq 10^6$ (if, e.g., $E_0 = 1 \text{ PeV}$)

$$N_{\text{EM}} \propto E_0^b \quad \text{with} \quad b = 1 + \frac{1-\beta}{10^{6(1-\beta)} - 1} \simeq 1.02. \quad (14.12)$$

The *superposition model* states that a primary nucleus with mass A and energy E_0 acts as A independent nucleons of energy E_0/A ; according to it, for the superposition of A nucleon showers (● Eq. 14.10) gives

$$X_{\max}^A = X_{\max}^p - X_R \ln A \quad (14.13)$$

$$N_{\mu}^A \simeq A \left(\frac{E_0/A}{E_c^{\pi}} \right)^{\beta} = A^{1-\beta} N_{\mu}. \quad (14.14)$$

It follows from (14.13) and (14.14) that the depth of shower maximum and the number of muons depend on the mass of the primary particle: showers originated by a proton develop lower in atmosphere (higher X_{\max}), while the higher the primary mass (at a given energy), the more muons are expected. Being the superposition of A nucleon subshowers, heavy nuclei showers will also have smaller shower to shower fluctuations as compared to protons. Gamma ray showers fluctuate much less and are muon poor, due to the small cross sections for meson production and muon pair creation. However, the superposition assumption is a simplification of the correct treatment of nucleus-nucleus interactions, which does not take into account the fact that in most collisions the number of interacting nucleons is not equal to that of the projectile. As shown in Ivanov (2010), it could result in an overestimation of the primary energy when evaluated using surface arrays particle density measurements.

The muon lateral distribution was parameterized first by Greisen (1960). Simplified parameterizations, tuned for each particular experiment, can be found in the literature. For example, Khristiansen et al. (1977):

$$\rho_{\mu}(r) \propto r^{-\alpha} \exp\left(\frac{-r}{r_M}\right). \quad (14.15)$$

As an example, the longitudinal and lateral distributions are shown in Fig. 14-16 for 10^{15} eV and different components.

3.1.3 Cherenkov Light

A charged particle of the air shower produces Cherenkov light if its velocity is such that $v/c > n$, where $n(z)$ is the index of refraction as a function of height z . Most of the Cherenkov light is emitted in air by electrons, for which the threshold energy is about 21 MeV at sea level (to be compared to about 500 MeV for muons), in a cone of emission which, well above threshold, is $\theta_C(z) = \cos^{-1}(1/\beta n(z))$, a decreasing function of the atmospheric height. The fraction of light F_C reaching the ground if emitted at an atmospheric depth x with zenith angle θ is (Hillas 1982)

$$F_C = \exp[-(1,020 - x)\sec \theta/\Lambda], \quad (14.16)$$

where Λ is the absorption length in g cm^{-2} . This light is in fact attenuated (in UV/blue) by scattering on molecules (Rayleigh scattering) and aerosol particles (Mie scattering), by scattering on water vapor in clouds or by the absorption by ozone molecules.

The light emitted by a typical gamma ray shower at 10 km results at ground in a ring focused at $\simeq 120$ m from the core; the number of emitted photons is about 0.1 cm at sea level.

The lateral density distribution of the Cherenkov light at ground is mainly determined by the Cherenkov angle and the Coulomb scattering and can be parameterized as (Fowler 2001)

$$C(r) = \begin{cases} C_{\text{crit}} e^{s(r_{\text{crit}}-r)} & \text{if } 30 \text{ m} < r \leq r_{\text{crit}}, \\ C_{\text{crit}} (r/r_{\text{crit}})^{-\beta} & \text{if } r_{\text{crit}} < r \leq 350 \text{ m}. \end{cases} \quad (14.17)$$

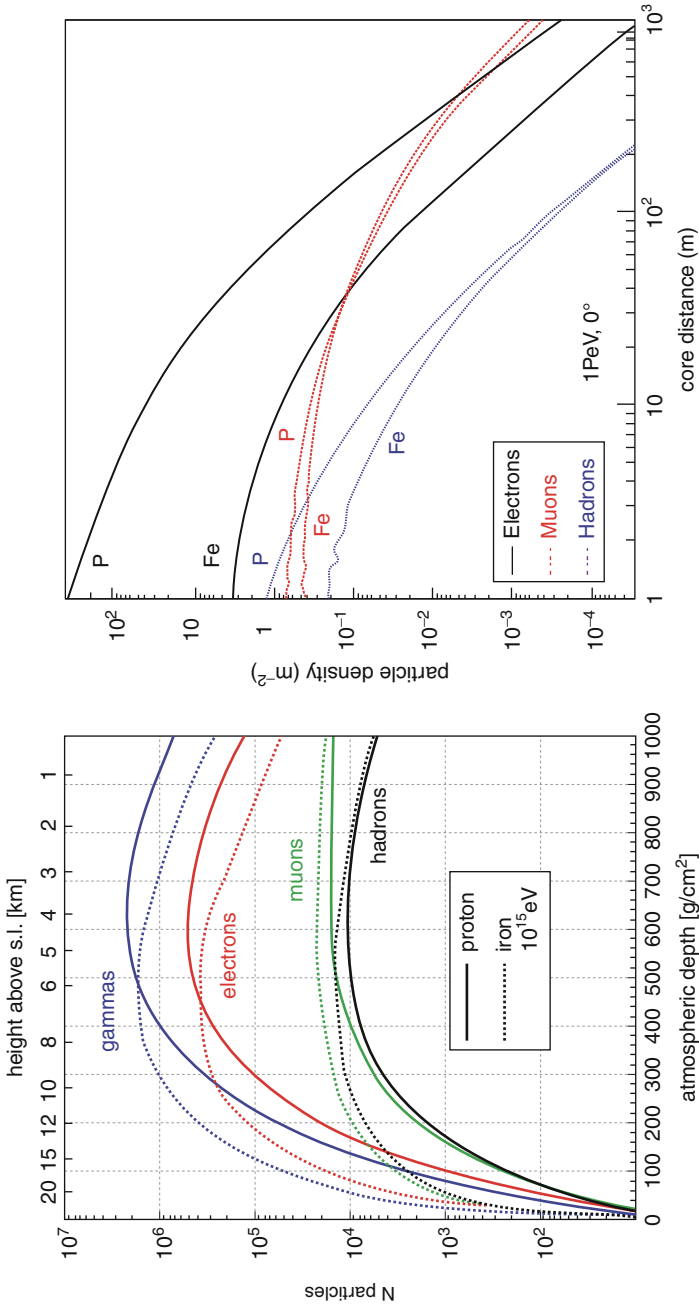


Fig. 14-16

Left: Longitudinal EAS development of the average total intensities (sizes). Right: Lateral distributions of the particle densities at sea level (From CORSIKA simulations (Heck et al. 1998))

Here, C_{crit} is the Cherenkov light at r_{crit} m from the core; the distance r_{crit} marks a sharp change of slope in the light distribution and is $\simeq 120$ m at sea level. s and β are the inner and outer slopes of the distribution.

A different contribution comes from the emission of direct Cherenkov light by primary particles with velocities above threshold before their first interaction in atmosphere. Being created higher in atmosphere, it generates on ground a light cone which is narrower (within $\simeq 100$ m) compared to that produced by the secondaries of the EAS. Since the intensity of this light is almost constant above a given energy, while the EAS one increases almost linearly with energy, the direct light is outshined by the EAS one above few hundreds TeV.

3.1.4 Fluorescence Light

Charged secondary particles of the air showers, mostly electrons and positrons, excite nitrogen molecules in the atmosphere: the de-excitation results in the isotropic emission of a fluorescence spectrum in the near UV region (300–400 nm). The fluorescence efficiency of photons, that is the ratio of the energy emitted in fluorescence by the excited gas to the energy deposited by the charged particles, is very low, of the order of 5×10^{-5} . This is, the reason why only at the highest energies, above $\simeq 10^{18}$ eV, the enormous number of particles allows this light to be detected. Quenching effects cause part of this energy to go to other molecules through collisions; since a shower typically crosses many km of altitude and since the collision rate depends on the average separation distance and velocity of the molecules, the fluorescence emission will depend on the gas pressure and temperature.

The number of emitted fluorescence photons can be written as

$$\frac{d^2 N_\gamma}{dX d\lambda} = Y(\lambda, P, T, e_\nu) \cdot \frac{dE_{\text{tot}}^{\text{dep}}}{dX}, \quad (14.18)$$

where $dE_{\text{tot}}^{\text{dep}}/dX$ is the energy deposited in the atmospheric depth dX . The fluorescence yield $Y(\lambda, P, T, e)$ depends on the wavelength λ , the air pressure and temperature P and T , and the vapor pressure e_ν . Its value is (5.05 ± 0.71) photons/MeV of energy deposited, in air at 293°K and 1,013 hPa in the 337 nm band according to Nagano et al. (2004). Exploiting the fact that the fluorescence light is mainly induced by MeV electrons, the emission mechanism and the absolute fluorescence yield have been studied in various laboratory experiments. Recent results are reviewed by Arqueros et al. (2008).

3.1.5 Radio Emission

Air showers generate coherent radio emission, first observed by Jelley et al. (1965), due to the cascade electrons emitting synchrotron radiation in the Earth magnetic field. In fact, these electrons, with $\langle E_e \rangle \simeq 30$ MeV, are spread in a thin shower front of ≤ 2 m, smaller than 1 wavelength for emitted radio pulses of 100 MHz. Up to this frequency, coherent emission can be expected. The electric field strength is proportional to the primary energy of the cosmic-ray particle initiating the shower.

Radio signals can arise also from coherent radio Cherenkov emission (the Askaryan effect): cascade e^+ can annihilate with the e^- of the medium (air), thus generating a $\simeq 20\%$ electron excess in the air shower, which behaves as a relativistic charge, emitting Cherenkov radiation

(Askaryan 1962). The emission in this case is coherent up to GHz in sufficiently dense materials like ice and can be exploited to reach the huge detection acceptance needed to study shower neutrinos.

3.2 Experimental Methods for Indirect Measurements

3.2.1 The Measure of the Charged Component

A classical air shower experiment (two examples are shown in [Fig. 14-17](#)) consists of an array of detectors, either scintillator counters or water Cherenkov tanks, distributed over a wide area, which surface is chosen depending on the rate of events to be studied. The separation between the detectors is tuned to match the scale of the shower footprint at the observation level (tens of meters in the PeV region, hundreds of meters to kilometers for the arrays studying the extreme energy region). At each location, the particle density of one or more charged components is measured with detectors of size suitable for the component under study (few m^2 for the electromagnetic one, much larger for muons and hadrons), together with the arrival times of the particles and their time spread. Due to the large number of secondary particles, the active area to be covered can be much smaller than the total one: sensitive/enclosed area ratio can go from $\approx 3 \times 10^{-3}$ for the EAS-TOP array ($37 \times 10 \text{ m}^2$ detectors over 10^5 m^2) to $\approx 5 \times 10^{-6}$ for the Pierre Auger Observatory ($1,600 \times 10 \text{ m}^2$ detectors over $3,000 \text{ km}^2$). This ratio, together with the altitude of the detector (i.e., thickness of the atmosphere above it) and the ability to detect different components of an EAS, determines the energy thresholds of different detectors. To lower the energy threshold, completely different apparatuses have also been designed, where the active area reaches at least 50% of the total one, thus allowing for the so-called full coverage (Aielli et al. 2009; Atkins et al. 2004).

Besides the aim of lowering the energy threshold, high-altitude experiments can be foreseen to study the air showers of higher energies at an early stage of development, to measure the direct primary spectrum of protons (before their interaction in the atmosphere), and to study hadronic production after few interaction lengths, that is, in the very forward region.



Fig. 14-17

The EAS-TOP, *left* (2,000 m a.s.l.), and KASCADE, *right* (sea level), air shower arrays (Aglietta et al. 1989; Klages et al. 1997)

The more penetrating muon component is generally measured by shielded detectors, like scintillator slabs: a shielding of thickness some radiation lengths (e.g., $20 X_0$ for KASCADE) can absorb the electromagnetic component without significantly affecting the muon one. Alternatively, one can use tracking devices (limited streamer or proportional tubes) or measure the muons, together with the electromagnetic component, in water Cherenkov tanks. In this last case, the electromagnetic particles are completely absorbed in water, while the muon signal is proportional to the track length in the detector. Since low-energy muons (below 1 GeV) mainly decay before reaching the ground, this component basically consists of muons with energies of few GeV. High-energy muons, with energies above few TeV, on the other hand, give information on the first interactions of the primary particle. They can be detected in underground laboratories, shielded by rock, water, or ice (Achterberg et al. 2006; Ahlen et al. 1993), either as single muons or bundles.

The core location and the total number of charged particles are obtained by means of a fit to a function describing their measured lateral distribution (e.g., (♣ Eq. 14.9) and (♣ Eq. 14.15)). The shower size (recall (♣ Eq. 14.8)) is typically evaluated above $N_e > 10^5$ with an accuracy $\sigma_{N_e}/N_e = 10\%$; the core position is determined within few meters (Aglietta et al. 1993). The arrival direction of the primary particle is derived from the measure of the arrival time of particles on the stations (the shower front). Most detectors have time resolution from 0.5 to few ns, with angular resolution typically below 0.5° , which can be evaluated from internal consistency of data. An absolute measurement of the angular resolution for an EAS array is possible by detecting the reduction in cosmic-ray intensity due to the “shadow” cast by the Moon and the Sun on the high-energy primary cosmic-ray flux. Being the Sun larger and much further away with respect to the Moon, they have basically the same angular diameter of $\approx 0.52^\circ$: the measure is in principle possible for arrays with angular resolution $\leq 1^\circ$ (Aglietta et al. 1991 and references therein). A very large sample of events is however needed to get a statistically significant result, due to the smallness of the effect.

Hadrons can be detected by means of calorimeters, measuring the energy dissipated by the incoming particles: two big devices have been used, for example, in the EAS-TOP (Adinolfi-Falcone et al. 1999) and KASCADE (Engler et al. 1999) experiments. The observables are in this case the hadron number and their energy sum.

3.2.2 The Measure of the Cherenkov Light

To observe the Cherenkov light emitted by the shower particles in the atmosphere, two techniques can be used:

- *Light-integrating detectors* combine a large angular acceptance with the advantages of Cherenkov light detection. They are used to measure the lateral distribution of the Cherenkov light with a grid of photomultipliers distributed over a large area on the ground, each enclosed in a Winston cone⁵ to help the light collection. This distribution is strongly related to the shower energy ($C_{120} \propto E^{1.07}$, see also (♣ Eq. 14.17)). The critical radius (here 120 m) is far enough from the core, at these energies, to minimize the fluctuations, and close enough to ensure a measurable light density. The dependence on the primary mass is fully included in the inner exponential slope s of the distribution, which is in fact a function

⁵Non-imaging light-collection devices with a parabolic shape and a reflective inner surface. Winston cones are often used to concentrate light from a large area onto a smaller photodetector or photomultiplier.

of the depth of the shower maximum. After many early attempts in the years 1972–1989, this technique was widely used in the dedicated HEGRA/AIROBICC experiment (Arqueros et al. 2000; Lorenz 1996) and subsequently in BLANCA (Fowler et al. 2001), both detectors operating in coincidence with a classical shower array. Another wide-angle Cherenkov array, TUNKA, is installed near lake Baikal, in Siberia (Chernov et al. 2005).

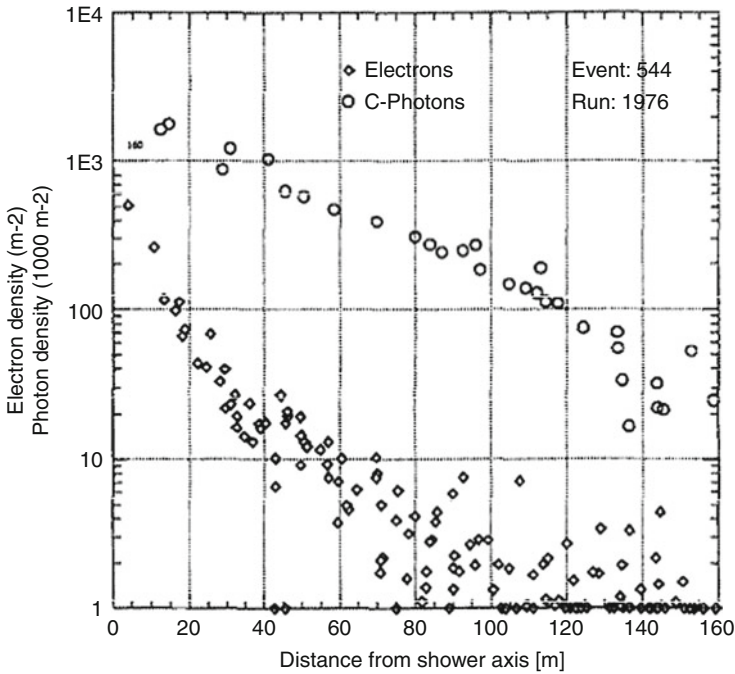
- *Imaging detectors* are used to reconstruct air showers generated by gamma ray primaries in TeV γ -astronomy, but the technique can also be used to study hadronic showers (Larsen et al. 2001). Basically, an image corresponding to the intensity pattern and direction of Cherenkov light is produced in the focal plane. When the direction of the shower and the distance of its core from the telescope are known, a simple geometrical procedure allows to measure the light emitted at each atmospheric depth. Its integral over the crossed atmospheric depth is used to determine the shower size, while a fit to the shape of the shower in the telescopes allows to derive the depth of maximum development X_{\max} in a way which is almost independent on simulations. The interpretation of the results in term of composition is however possible only with the help of Monte Carlo simulations. Imaging detectors can also be used to detect the direct Cherenkov light emitted by the primary particle before interacting. Since this light is emitted in a cone with emission angle between 0.15° and 0.3° , telescopes equipped with cameras using $\leq 0.1^\circ$ pixel size are needed. The light can be seen (Aharonian et al. 2007) as a single high-intensity pixel between the reconstructed shower direction and the center of gravity of the EAS image.

The broader lateral distribution (see [Fig. 14-18](#)), due to the smaller absorption of photons in atmosphere, and the high photon density, which means a better signal-to-noise ratio even for smaller arrays, are the main advantages in using these detectors as compared to classical EAS arrays for charged particle. On the other hand, the duty cycle for Cherenkov observations does not go above 10%, since they are possible only in clear moonless nights.

3.2.3 The Measure of the Fluorescence

The fluorescence light is collected with mirrors and projected onto a camera generally made of a large number of photomultiplier tubes (pixels), which record a time sequence of light. The EAS appears as a trace of illuminated pixels. As simulations show, most of the total energy of the shower is detectable as ionization energy, but a correction must be introduced to take into account the fraction of energy not contributing to the total signal (carried below observation level in the ground, by muons and neutrinos). This fraction decreases with increasing energy because at higher energy pions mostly interact, producing π^0 s of lower energy which in turn go to feed the electromagnetic component. At energies around 1 EeV, $\simeq 15\%$ and $\simeq 10\%$ of the energy is missing for iron and proton primaries, respectively, but the dependence on the primary mass becomes very small as energy increases. This quantity is also mildly dependent on the interaction model employed in the Monte Carlo, but being the energy mostly released in the electromagnetic component, this reliance remains at the level of few percent.

The geometry of the shower, together with the evaluation of the Cherenkov light background and of the atmospheric absorption, gives a complete reconstruction of the event. The fluorescence telescopes can be used in monocular (one telescope) or in stereo (two or more telescopes) mode. In the latter case, the angular resolution can reach 0.6° .



■ Fig. 14-18

Cherenkov photon density as measured by HEGRA-AIROBICC (Karle et al. 1995), compared to the electron density at ground level by the scintillator array

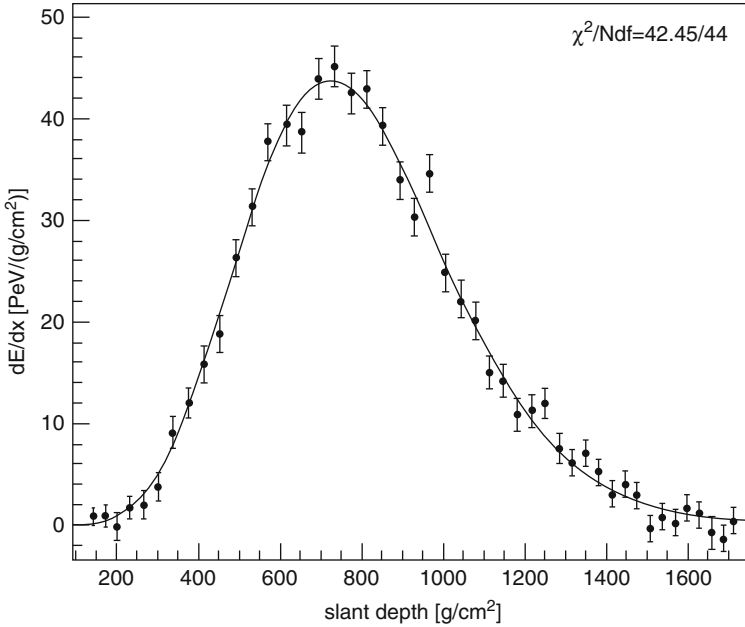
This technique was firstly used by the Fly's Eye experiment (Baltrusaitis et al. 1999); 8 years of data came from its successor HiRes (Boyer et al. 2002).

In the Pierre Auger Observatory, 24 fluorescence telescopes are employed (Abraham et al. 2010a), located in four stations overlooking the surface array. The hybrid technique, combining data recorded by both the surface and the fluorescence detectors, has great advantages: (a) the energy scale is set with the fluorescence telescopes and is thus nearly independent of shower simulations; (b) the shower arrival directions are determined with very high precision, cross-checking the angular resolution derived from the surface stations; (c) the two techniques are complementary, and different observables are measured in a redundant way so that many cross-checks are possible.

An example showing the energy deposit in atmosphere by a EAS as reconstructed in the Pierre Auger fluorescence detector is shown in ● Fig. 14-19.

3.2.4 Energy and Composition Estimators

The experimental observables which are measured in order to extract information about the energy spectrum are the charged components of showers as measured by ground-based detectors with scintillator counters, muon, and hadron detectors, the Cherenkov light produced by



■ Fig. 14-19

Reconstructed energy deposit profile for an EAS in the Pierre Auger fluorescence detector. The reconstructed energy is $(3.0 \pm 0.2) 10^{19}$ eV (Abraham et al. 2010a)

shower particles as they propagate through the atmosphere, and at higher energy, the fluorescence emission. Mass and energy of the primary particles are strictly correlated and, in general, the EAS development depends on the interplay of the two.

In most cases a single observable, e.g., the electromagnetic or muonic shower size (i.e., the total number of electromagnetic charged particles N_e or the total number of muons N_μ), is used to derive the all-particle energy spectrum. The relation of the size to the energy is obtained from simulations with some hypotheses on the mean primary mass. As an example, Aglietta et al. (1999) find the relation between shower size and primary energy:

$$N_e(E_0, A) = \alpha (A_{\text{eff}}) E^{\beta(A_{\text{eff}})}, \quad (14.19)$$

where the normalization α and the slope β depend on the effective mass A_{eff} , calculated from the extrapolation of the spectra $\Phi_i(N_e)$ of the individual elements (each with atomic mass A_i) measured at low energies by direct measurements:

$$A_{\text{eff}}(N_e) = \frac{\sum_i A_i \Phi_i(N_e)}{\sum_i \Phi_i(N_e)}.$$

Above the knee, a rigidity dependent cutoff is used.

Combining the measure of the electromagnetic and muonic sizes, it is possible to obtain an energy determination almost independently of the primary mass; this method was exploited, for example, in the CASA-MIA experiment (Glasmacher et al. 1999), where

$$E_0 = 0.8 \text{ GeV} (N_e + 25 N_\mu) \quad (14.20)$$

for any primary mass within an uncertainty of $\approx 5\%$.

An estimate of energy can be derived from a combination of the amount of Cherenkov light and the location of shower maximum; the lateral distribution and intensity of Cherenkov light at a given total energy depend in fact both on the primary particle mass, hence on the mean X_{\max} , and the distance of the measurement from the shower.

For giant arrays, like those measuring EAS above the knee region up to the highest energies, the energy is generally determined by measuring the particle densities at a specific distance from the core. In fact, the effects of intrinsic shower to shower fluctuations are minimized if the signal is measured at an optimal core distance, which depends only on the detector spacing in the array (Hillas et al. 1971). Following this work, the particle density $S(600)$ at 600 m distance was used as energy estimator in the Haverah Park experiment and later on in the AGASA array (Takeda et al. 2003a), while in the Pierre Auger experiment (Abraham et al. 2008), where detectors are spaced on a 1.5 km grid, this distance is increased to 1 km. Finally, for fluorescence telescopes, the energy is measured almost calorimetrically by integrating the fluorescence light along the shower path.

The deconvolution of the primary energy and mass can be successfully performed in surface detectors by correlating different observables, for example, N_e and N_μ . The procedure consists basically of an unfolding of the two-dimensional $N_e - N_\mu$ distribution of the electromagnetic and muonic sizes of the EAS into the energy spectra of the primary mass groups, their correlation thus being taken into account. The number of events in each cell $(N_e, N_\mu)^j$ can be considered as the superposition of contributions from different primary particles of mass A and energy E :

$$N_j = S_s T_m \sum_{A=1}^{N_A} \int_{\Omega} \int_{-\infty}^{+\infty} \frac{dJ_A}{d\log E} \times p_A d\log E d\Omega, \quad (14.21)$$

where $dJ_A/d\log E$ is the differential flux of an element with mass number A and the summation is carried out for all elements present in the primary cosmic radiation. T_m represents the measurement time over a sampling area S_s ; $d\Omega = \sin\theta d\theta d\phi$ is the differential solid angle. The probability p_A to measure at ground the sizes N_e, N_μ from a shower of primary energy E and primary mass is evaluated using Monte Carlo simulations and includes the shower fluctuations, the detection and reconstruction efficiencies (Antoni et al. 2005).

Techniques able to extract an almost pure light component (p +He) without the need for deconvolution have been used in Aglietta et al. (2004), exploiting the space correlation of the EAS-TOP surface detector and the MACRO underground apparatus. The two experiments were separated by a rock thickness ranging from 1,100 up to 1,300 m depending on the angle and located at a respective zenith angle of about 30° . The muon energy threshold at the surface for muons reaching the MACRO depth ranged between 1.3 and 1.8 TeV within the effective area of EAS-TOP. Light primaries can be selected based on their energy/nucleon by means of high-energy muons, while the associated Cherenkov light detected on surface is proportional to the primary energy.

Air shower arrays with large area muon detectors can study different regions of the muon multiplicity distribution as a function of the electron component of showers, with high sensitivity to different nuclear groups present in the primary flux (Gupta et al. 2009).

Detectors measuring the Cherenkov or fluorescence light produced by EAS in the atmosphere can derive the primary mass by measuring the depth of maximum development of the showers, X_{\max} (see (🔗 Eq. 14.13)).

3.2.5 The Link to Particle Physics

The interpretation of the ground-level observations in terms of primary particle characteristics is far from straightforward, because of the significant (mass dependent) fluctuations of EAS observables and of the strong dependence on hadronic interaction models used in the simulations of the production and propagation of particles through the atmosphere. At present, the latter constitutes the dominant source of systematic uncertainty.

Most of the observables that are relevant for the shower development (the total inelastic cross sections, the multiplicities of the final states, and the inclusive energy spectra) must be obtained from extrapolations of the measurements in accelerator experiments, which are performed in a much lower energy range. The new measurements at the CERN LHC collider will reach $\sqrt{s} = 14$ TeV, that is, $E_{\text{lab}} \approx 10^{17}$ eV, well below the maximum energy observed in cosmic rays, and only for p - p interactions, while Pb-Pb collisions at $\sqrt{s} = 5.5$ TeV n^{-1} are foreseen. Furthermore, the kinematic region of interest for cosmic-ray interactions is the projectile fragmentation region, while in collider experiments, the central region of the interaction is best explored. A couple of experiments starting at LHC will be of particular interest for cosmic-ray physicists: LHCf (Adriani et al. 2008) will explore the very forward region of the interaction at $|\eta| > 8.5$,⁶ and TOTEM (Antchev et al. 2010) will derive the p - p cross section in the range $3.1 \leq |\eta| \leq 6.5$.


On the other hand, EAS measurements are the only way to give information about very high-energy hadronic interactions, and can perform precise tests to constrain the models. The data from KASCADE, for example, allowed to compare the measured correlations between different hadronic observables (such as number of hadrons, their energy sum, their lateral distribution, and energy spectra) with the number of muons and electrons at ground level with the results of the CORSIKA simulation code for various models (Apel et al. 2007 and references therein).

The measurement of the inelastic proton-air cross section can be performed using EAS arrays from the mean depth of the first interaction and its fluctuations (Ulrich et al. (2009) and references therein). They can be determined indirectly either investigating the unaccompanied hadrons, at low energies, or basing on the analysis of the muon and electromagnetic shower sizes using simulations.

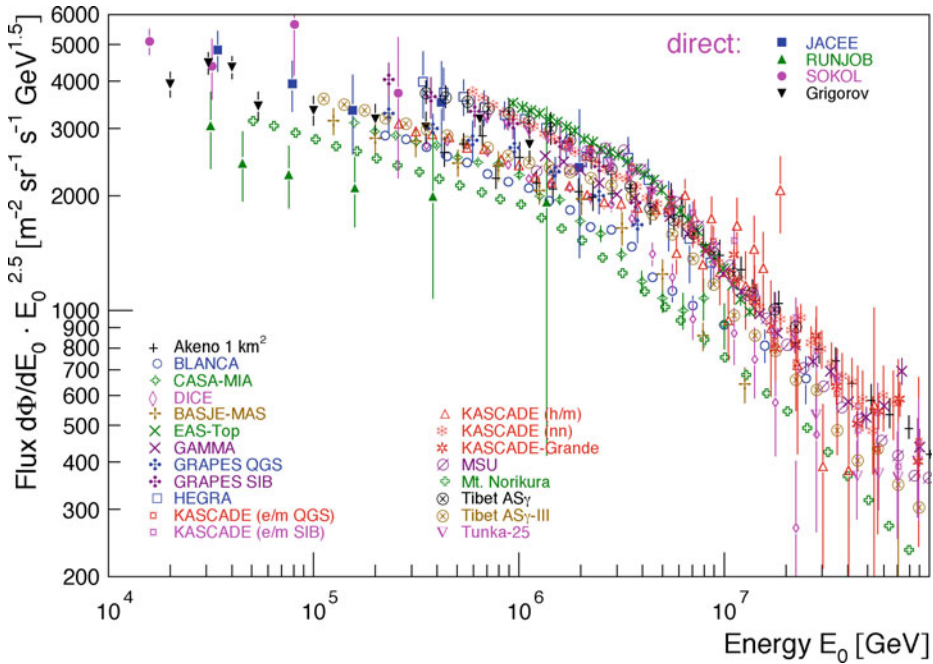
Most recently, the Pierre Auger Collaboration used their hybrid data to analyze the shape of the depth of shower maximum distribution: the choice of showers more deeply penetrating in the atmosphere allows to obtain a proton-rich sample of events (Ulrich et al. 2011). The result favors a moderately slow rise of the cross-section towards higher energies, a result that seems to be confirmed in the first data from LHC (Aad et al. 2011).

3.3 The Knee Region

3.3.1 The Energy Spectrum

A huge number of experiments contributed to the measure of the all-particle energy spectrum, shown in  Fig. 14-1. In the lowest energy region, the fluxes are obtained by means of direct

⁶The pseudorapidity $\eta = -\ln[\tan(\theta/2)]$ measures the angle of the particle with respect to the beam direction. The forward region is characterized by $|\eta| > 4$



■ Fig. 14-20

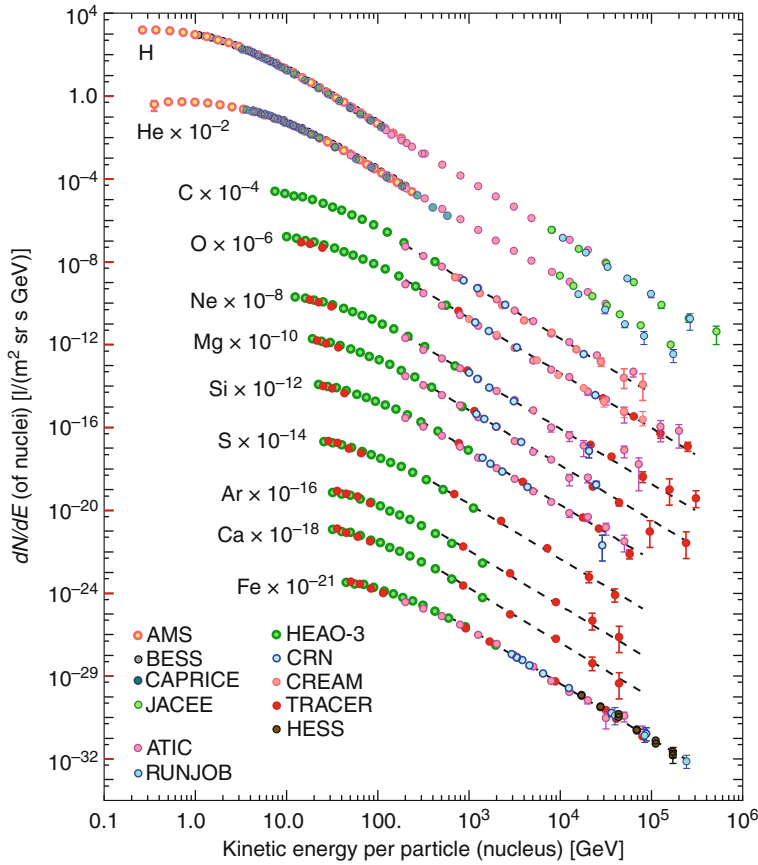
The all-particle energy spectrum in the knee region (Blümer et al. 2009)

measurements above the atmosphere, the RunJob (Derbina et al. 2005) and JACEE (Asakimori et al. 1998) being the only two balloon experiments extending above 100 TeV. The all-particle spectrum above 10^{13} eV is shown in Fig. 14-20. All experiments agree on a spectrum described by a power law with a slope changing from $\langle \gamma \rangle = (2.66 \pm 0.06)$ below the knee to $\langle \gamma \rangle = (3.10 \pm 0.09)$ above. The mean value of the knee energy is $E_{\text{knee}} = (3.9 \pm 0.7)$ PeV.

In the lower energy region, the results from balloons and EAS experiments overlap, showing a good agreement considering that indirect data below about 10^{15} eV are dominated by systematic uncertainties while direct measurements suffer from statistical ones.

The systematic uncertainties depend on the fact that the all-particle energy spectrum has been obtained using different observables, dissimilar assumptions on composition, or in some case no assumption at all, in apparatuses located at quite different altitudes (from underground laboratories to the sea level to high mountains). Each experiment employs a specific “particle” unit to define the signal, and the observables are measured with quite dissimilar accuracies. Furthermore, the methods of conversion from the observables to the energy are often based on different simulation codes which in turn exploit various hadronic interaction codes. A recent discussion of the differences among the experiments can be found in Nagano (2009).

More information can be gathered by measuring the spectra of single elements: a recent compilation of their spectra is shown in Fig. 14-21 from the results of direct measurements. Updated data can be found in Ahn et al. (2010), Obermeier et al. (2011), Mewaldt et al. (2010), and Rauch et al. (2009).



■ Fig. 14-21

Single elements spectra from direct measurements (Nakamura et al. 2010)

In addition to the balloon and satellite results, direct measurements of single elements can be obtained by measuring the Cherenkov light emitted directly by the primary particle in the $\approx 8\text{--}15\text{ g cm}^{-2}$ of atmosphere above the first interaction with an atmospheric nucleus. Exploiting the different emission angle and time of arrival of the direct and shower Cherenkov contributions (see ▶ Sect. 3.2), it is possible to separate the two signals. In this way, the energy spectrum of iron nuclei was determined between 13 and 200 TeV employing the HESS imaging atmospheric Cherenkov telescopes (Aharonian et al. 2007).

The knowledge of the energy spectra of single (or groups) of elements from direct measurements has been extended to higher energies by exploiting the ability of the modern ground experiments to detect the different components of the EAS. A compilation of the world results is shown in ▶ Figs. 14-22 and ▶ 14-23 for the single (or groups of) element spectra (Bertaino et al. 2008).

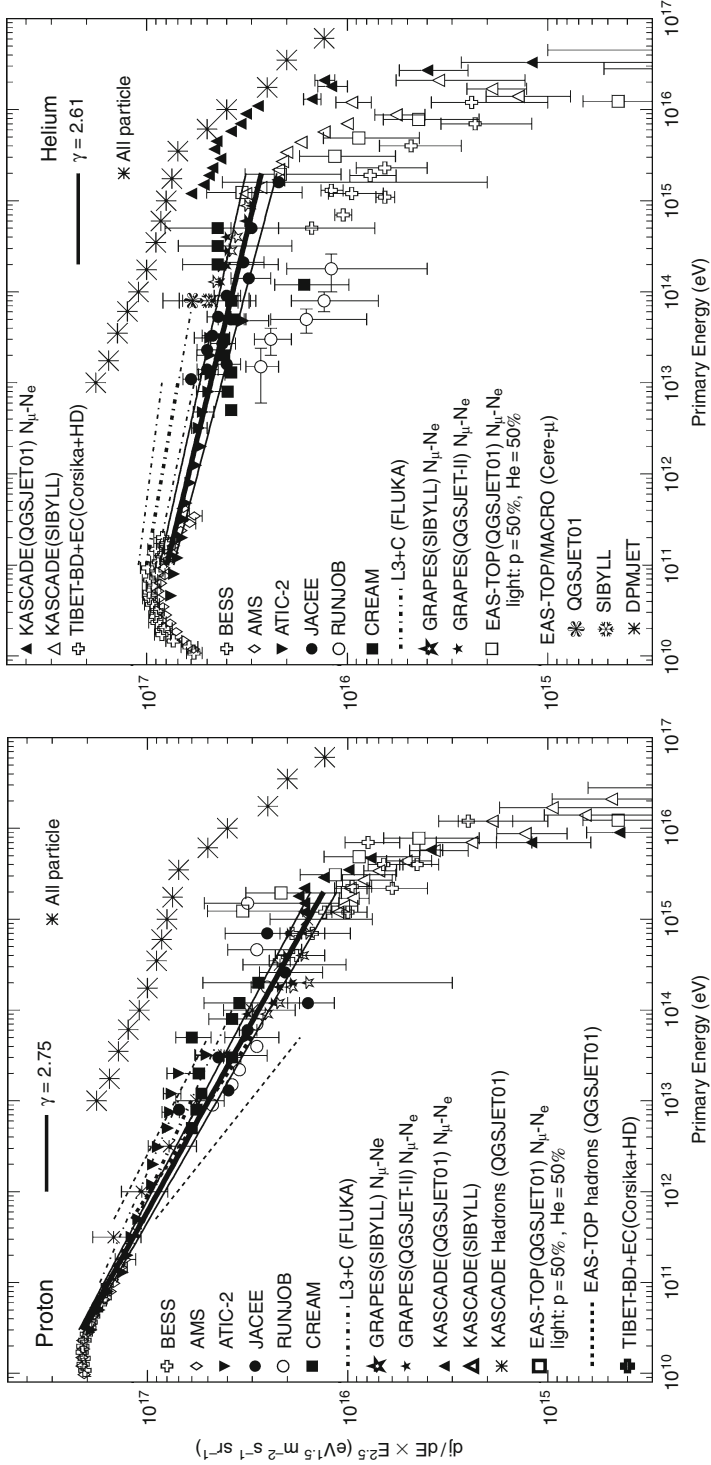
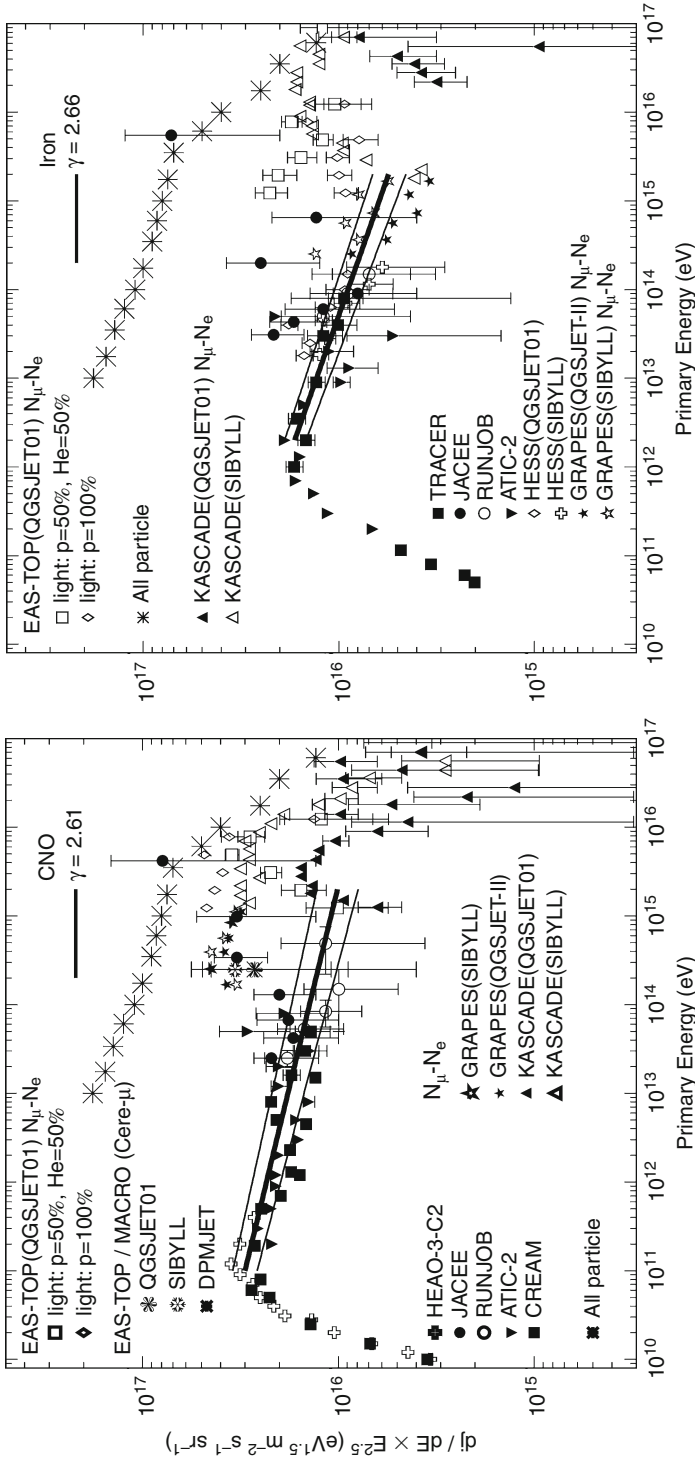


Fig. 14-22 Compilation of the proton and helium spectra (Bertina et al. 2008). The all-particle spectrum is shown by asterisks for reference



■ Fig. 14-23
 Compilation of the CNO and iron spectra (Bertina et al. 2008). The all-particle spectrum is shown by asterisks for reference

The spectra of the different elements can be fitted using a power law up to the corresponding knee. For protons, for which there is a remarkable agreement of all techniques in the whole energy range

$$\frac{dN}{dE} = (8.8 \pm 0.5) E^{-2.75 \pm 0.01} \text{ m}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ TeV}^{-1}, \quad (14.22)$$

while for all the other elements, the slope is close to 2.61. Most experiments confirm a difference between the proton and helium slopes, which could be interpreted with two different types of sources/acceleration mechanisms for the two elements (Biermann 1993). The only contradicting results come from the RunJob collaboration; however, a very recent reanalysis of their data (Kopenkin and Sinzi 2009) attributes the differences as being due to the very low statistics and systematic uncertainties, especially in the high-energy region. The larger spread among the results for the CNO can be partly explained by the slightly different definitions of “CNO” group in the various experiments; some of the data shown in the Fe plot refer to a more general group of heavy elements around iron.

The measured spectra show that the knee in the all-particle spectrum is mainly due to the light elements suppression. The cutoffs in the different spectra seem to show up at energies proportional to the nuclear charge; the most recent results (Apel et al. 2009) confirm this conclusion, underlining once more the limiting factor due to the uncertainty in the hadronic interaction models.

Dividing their air shower events in electron-rich and electron-poor groups, based on the measure of the charged particle and muonic components, Apel et al. (2011) found the first experimental evidence for a knee-like break in the cosmic-ray spectrum of heavy primaries at about 9×10^{16} eV, in agreement with models where the components of the primary beam bend at subsequent knees proportional to the charge of the primary nuclei.

A different conclusion is reached in the Tibet experiment (Amenomori et al. 2011), where data indicate a heavy component dominance around the knee, with light nuclei bending below the all-particle knee and contributing to not more than 30% at the knee. According to the authors, the disagreement with other data (e.g., KASCADE) can be attributed to the different kinematic region explored and to model dependence. The result could be explained either assuming cosmic rays produced in nearby sources with source composition dominated by heavy nuclei or nonlinear effects in the diffusive shock acceleration mechanism, but the limited statistics could affect the conclusion.

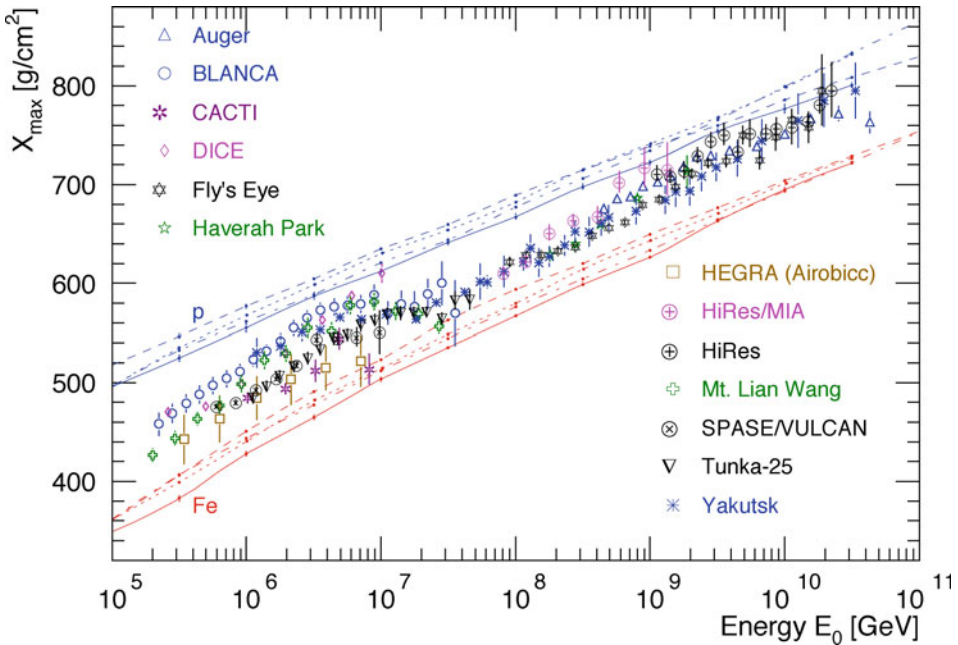
3.3.2 The Composition in the Knee Region

Above about 100 TeV, the particle fluxes are very low and the measurements are subject to large fluctuations; this is why the results obtained with the different techniques on the primary composition are generally given in terms of the mean logarithmic mass, defined as

$$\langle \ln A \rangle = \sum_i r_i \ln A_i, \quad (14.23)$$

where r_i is the relative fraction of nuclei of mass A_i . The experimental measurement of this quantity is performed exploiting either the proportionality of $\langle \ln A \rangle$ to the mean shower maximum or that of A to the ratio of the charged components of the shower. In fact, in the frame of a simple description of the shower development in atmosphere (see [Sect. 3.1](#)) and using the superposition model, ([Eq. 14.13](#)) gives the dependence of $\langle X_{\text{max}}^A \rangle$ on $\ln A$, while recalling ([Eqs. 14.12](#)) and ([14.14](#)), the ratio of the electromagnetic and muonic particle sizes

$$N_e/N_\mu \propto \left(\frac{E_0}{A} \right)^{0.15}. \quad (14.24)$$



■ Fig. 14-24

Average depth of the shower maximum as a function of energy as derived from Cherenkov or fluorescence detectors (Blümer et al. 2009)

The results on the measure of $\langle X_{\max} \rangle$ are shown in **Fig. 14-24**. Monte Carlo simulations with specific hadronic interaction model choice are used to get the expected average depths of maximum for protons and iron, $X_{\max}^{p,\text{sim}}$ and $X_{\max}^{\text{Fe},\text{sim}}$, so that

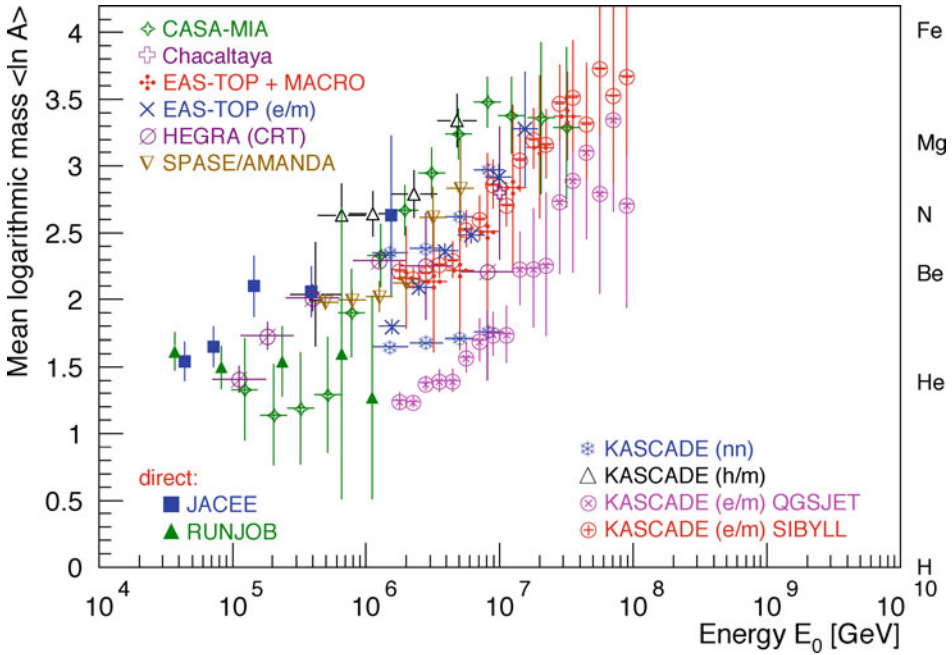
$$\langle \ln A \rangle = \frac{\langle X_{\max} \rangle - X_{\max}^{p,\text{sim}}}{(X_{\max}^{\text{Fe},\text{sim}} - X_{\max}^{p,\text{sim}})} \cdot \ln A_{\text{Fe}}. \quad (14.25)$$

The composition is getting lighter towards the knee, while the opposite happens above. From 10^7 GeV up to the transition region, it remains almost constant.

The mean logarithmic mass from the measurements of the charged components of the showers is shown in **Fig. 14-25**, obtained using the QGSJet-01 hadronic interaction model. The increase in $\langle \ln A \rangle$ across the knee region is clear, but we do not see it to become lighter towards the knee, as it was observed by means of the measurement of X_{\max} . A better agreement among the different methods can be obtained modifying some of the features of the models; for example, a decrease in the inelastic cross section and an increase in the elasticity turn into a deeper depth of shower maximum, which reflects into the number of particles produced at ground (Hoerandel et al. 2009).

3.4 Models for the Knee

Figures 14-22 and **14-23** show a compilation (Bertaina et al. 2008) of the data on the single (or groups of) element spectra; together with those from composition and anisotropy, these results can give some hints about the origin of the spectral feature of the knee.



■ Fig. 14-25

The mean logarithmic mass from the measurement of charged components of EAS at ground as a function of energy (Blümer et al. 2009)

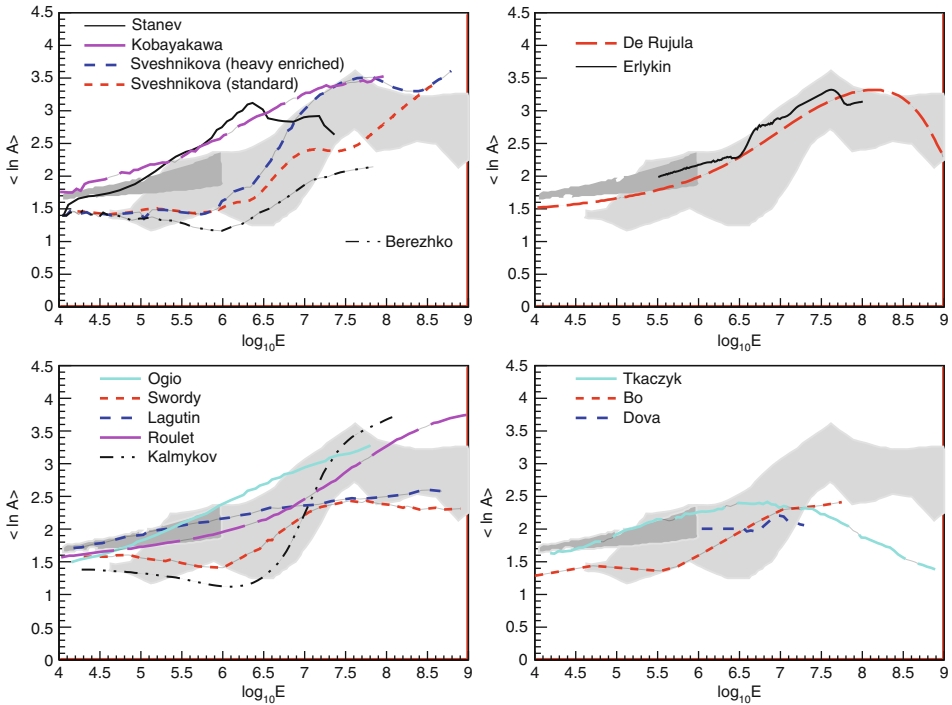
The most popular scenario attributes the knee to the maximum energy attainable by a relativistic particle of charge Ze and energy E (PeV) during the acceleration process in a site with magnetic field B and size L (see (Eq. 14.1)).

The actual value of the break depends on the properties of the source and of the medium, on the strength of the magnetic field, and its orientation with respect to the particles motion. While breaks around $E_{\max} \propto Z \times 10^{14}$ eV are predicted in (Biermann 1993; Kobayakawa et al. 2002; Stanev et al. 1993), higher energies can be reached if B is increased, for example, taking into account preacceleration in the precursor wind (Berezhko et al. 1999) or on the contrary increasing the shock parameters and summing the contributions of different kinds of supernovae (Sveshnikova et al. 2003).

In (Fig. 14-26 (top left)), the prediction of these models are compared to the experimental results obtained by means of direct (dark-gray area) and indirect (light-gray area) measurements.

Pertaining to the same group, the model of Erlykin and Wolfendale (2001) and Erlykin and Wolfendale (2009) assumes that, out of a background of many undefined sources, a single close-by SNR is responsible for the knee. From a recent comparison among the spectra measured by ten different groups (Erlykin and Wolfendale 2011), the knee appears sharper⁷ than that predicted in diffusion models, helium is the dominant nucleus around the knee, and a second peak around 50–80 PeV (maybe the iron peak) is found. The mean logarithmic mass derived

⁷The “sharpness” of the knee is here defined as $S = -d^2 \log I / d(\log E)^2$, where $I(E)$ is the primary cosmic-ray energy spectrum



■ Fig. 14-26

The mean logarithmic mass from direct (*dark gray*) and indirect (*light gray*) measurements as a function of the logarithm of energy, compared with different models. *Top left*: Acceleration models. *Top right*: Cannonball and single source models. *Bottom line*: Propagation in the Galaxy (*left*) and interaction with background particles (*right*) (see text for details) (Adapted and modified from Hoerandel (2004))

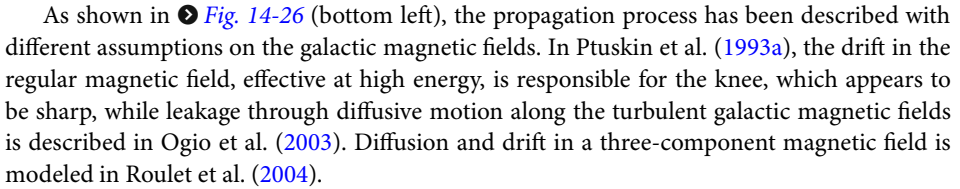
from this model is shown in ● Fig. 14-26 (top right); viable candidates for this single source are suggested to be B0656+14 in the Monogem ring or J0833–45 in the Vela SNR.

A single type of source, the “cannon ball,”⁸ is supposed to be the origin of all nonsolar cosmic rays, at all energies in Dar and De Rujula (2008): a sequence of magnetic collisions in the turbulent field of the cannonball accelerates cosmic rays to higher and higher energies. The model is developed as a generalization of the one used for gamma ray bursts (Dar and De Rujula 2004). The prediction is shown in the same figure, and it seems to be in good agreement with data. However, further work is needed in this model to understand the mechanisms of confinement and acceleration, and the diffusion to Earth must be proven negligible.

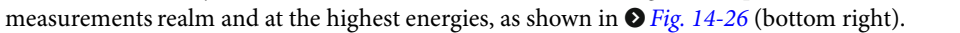
The knee can also be generated from propagation effects: in this case, as the energy of the cosmic rays increases, their confinement in the Galaxy becomes more difficult and their escape easier. Diffusive shock acceleration with energy dependent propagation path length has been used by Swordy (1995), with the introduction of a “residual path length” $\lambda_{\text{res}} \approx 0.013 \text{ g cm}^{-2}$

⁸Jet of plasmoids of ordinary matter emitted in the explosion of a core-collapse SN, in addition to the ejection of a nonrelativistic spherical shell.

which is supposed to provide a minimum path length at high energy: $\lambda_e = \lambda_0 \left(\frac{R}{r_M}\right)^{-\delta} + \lambda_{\text{res}}$. The propagation process has been described in terms of anomalous diffusion in Lagutin et al. (2001), where the knee appears to be due to the fractal structure of the galactic magnetic fields.

As shown in  Fig. 14-26 (bottom left), the propagation process has been described with different assumptions on the galactic magnetic fields. In Ptuskin et al. (1993a), the drift in the regular magnetic field, effective at high energy, is responsible for the knee, which appears to be sharp, while leakage through diffusive motion along the turbulent galactic magnetic fields is described in Ogio et al. (2003). Diffusion and drift in a three-component magnetic field is modeled in Roulet et al. (2004).

A combination of both acceleration and propagation effects could give an acceptable description of the data up to 10^{16} eV.

According to a third group of models, the knee is a threshold effect, due to the interactions of charged cosmic rays with background particles. Massive neutrinos, as proposed in Dova et al. (2001) and Wigmans (2003), could in principle explain the spectral features. Besides producing a too light composition above the knee, they appear to be excluded by the results of WMAP and 2dFGRS (Hannestad 1993). Photodisintegration of the cosmic-ray nuclei by interactions with optical and soft UV photons in the source region has also been considered (Candia et al. 2002; Karakula and Tkaczyk 2004), but it does not succeed in describing the composition in the direct measurements realm and at the highest energies, as shown in  Fig. 14-26 (bottom right).

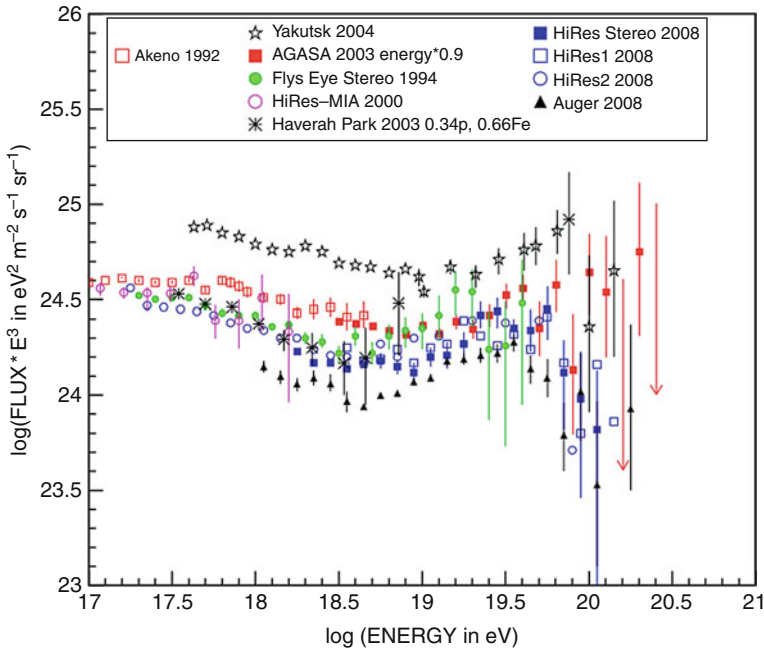
A more recent proposal relates the knee to the e^+e^- pair production by cosmic-ray nuclei interacting with background photons (Wang et al. 2010).

Finally, new hadronic interactions could take place in the atmosphere at very high energies. The knee would be in this case just due to an underestimation of energies of the shower particles. The missing energy could be transferred to unobservable particles, for example, gravitons (Kazanas and Nikolaidis 2001); a test of the model will come from LHC results. The production of PeV muons from the leptonic decay of hypothetical heavy short-lived particles was proposed in Petrukhin (2003). They would not be detected by EAS arrays (which generally count muons but do not measure their energy). An exotic component or high-energy atmospheric muons seems however to be already excluded by the measurements of Baikal ν telescope (Wischniewski et al. 2005).

4 Above 100 PeV n^{-1} : The Onset of Extragalactic Cosmic Rays

4.1 Energy Spectrum and Composition

Different experiments have collected data in the galactic/extragalactic transition region, from the second knee to the ankle: Haverah Park (Ave et al. 2003; Lawrence et al. 1991), Akeno (Nagano et al. 1992b), Fly's Eye (Bird et al. 1994), Yakutsk (Egorova et al. 2001), HiRes-MIA (Abu-Zayyad et al. 2001), AGASA (Takeda et al. 2003b), and Auger (Abraham et al. 2010b). Some of them are (or were) surface detectors: water Cherenkov detectors over an area of $\approx 12 \text{ km}^2$ at Haverah Park, and scintillators spread over 100 km^2 at AGASA. The Yakutsk array in Siberia covered an area of $\approx 18 \text{ km}^2$ with surface scintillators, underground detectors, and a photomultiplier system to measure the Cherenkov light in air. Fly's Eye and its successor HiRes are based on the fluorescence technique. The latter used a system of two detector stations 12.6 km apart, with 22 and 42 telescopes, respectively, each with 3.7-m diameter spherical



■ Fig. 14-27

Primary energy spectrum in the transition region and above. The differential flux in each bin is multiplied by an energy-dependent power E^3 (Nagano 2009)

mirrors. The Pierre Auger Observatory exploits both the surface array (1,600 water detectors over 3,000 km²) and the 24 fluorescence telescopes overlooking the apparatus from four sites to measure EAS in hybrid mode.

The primary energy spectrum is shown in **► Fig. 14-27**, multiplied by a factor E^3 to better show the deviations from a pure power law; the “ankle” feature is clearly visible in all data, at an energy corresponding to $\log_{10}(E/eV) = 18.6/18.65$.

The systematic uncertainties play a leading role here: they are entangled with the statistical ones in the spectrum representation of the figure (due to the E^3 multiplication), but it can be shown that a good agreement both in the flux normalization and in the ankle position can be reached if they are properly taken into account. For a surface detector like AGASA, they depend mainly on the simulations used to relate the signal measured at 600 m, $S(600)$, to energy. In the case of HiRes, the overall systematic uncertainty on the energy scale is $\simeq 17\%$ (Abbasi et al. 2004); in this apparatus, the aperture is rapidly growing with energy, since at higher energy the showers are brighter and can thus be detected at larger distances, and it is determined by simulations.

In the Pierre Auger Observatory, the energy-dependent exposure includes the trigger, reconstruction and selection efficiencies, and the evolution of the detector in time (through the construction phase). Its total systematic uncertainty is estimated to be 10% at 1 EeV, decreasing to 6% above 10 EeV. The uncertainties in the evaluation of the energy scale depend on the fluorescence yield, the absolute calibration, and the reconstruction method and amount to $\simeq 22\%$ (Abraham et al. 2008). Moreover, the hybrid technique exploited in the Auger experiment allows to directly correlate the signal measured by the ground array at 1,000 m from the shower axis,

$S(1,000)$, to the calorimetric energy measured by the fluorescence detector. In such a way, the energy assignment is almost independent of simulations, which enter only in the evaluation of the energy which can be missed because it is carried by neutrinos and muons and thus does not contribute to the fluorescence signal.

Our knowledge of the mass of primary cosmic rays in the transition region and above is quite poor. The conclusions reached by the different groups depend on the methods used to measure the mass, and most importantly, they are dependent, to a greater or lesser extent, upon the interaction model that is assumed in the simulations needed to interpret the data.

The composition can be studied by measuring the depth of maximum development of showers, X_{\max} . The results from Fly's Eye (Bird et al. 1994), HiRes-MIA (Abu-Zayyad et al. 2001), Hires-stereo (Abbasi et al. 2005), and Auger (Abraham et al. 2010c) are shown in Fig. 14-28, together with the prediction of different models for proton or iron primaries.

As it is clear from the figure, the differences among the predictions are quite big; the model ambiguity is even higher when using the observables from surface arrays since the predictions for the number of muons at ground differ as much as 30%. However, if their description of hadronic interactions at these energies is realistic, then a trend from heavier to lighter composition is observed in data from Fly's Eye and HiRes, while a more mixed composition is suggested by the Auger data. Above 10 EeV, the latter suggest a gradual increase of the primary cosmic-ray mass, a conclusion also confirmed by the analysis of the magnitude of fluctuation of the position of X_{\max} , which, for fixed primary energy, is expected to be larger for protons than for iron nuclei.

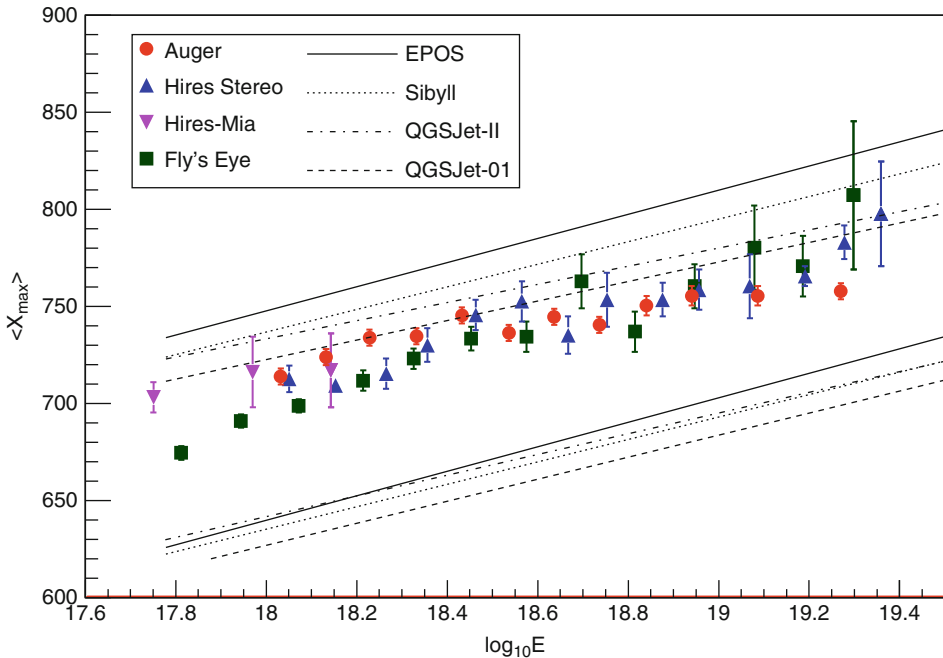


Fig. 14-28

Mean depth of shower maximum vs. primary energy, compared to different hadronic interaction models for protons and iron (Modified from Abraham et al. (2010c))


4.2 The Astrophysical Interpretation of the Transition

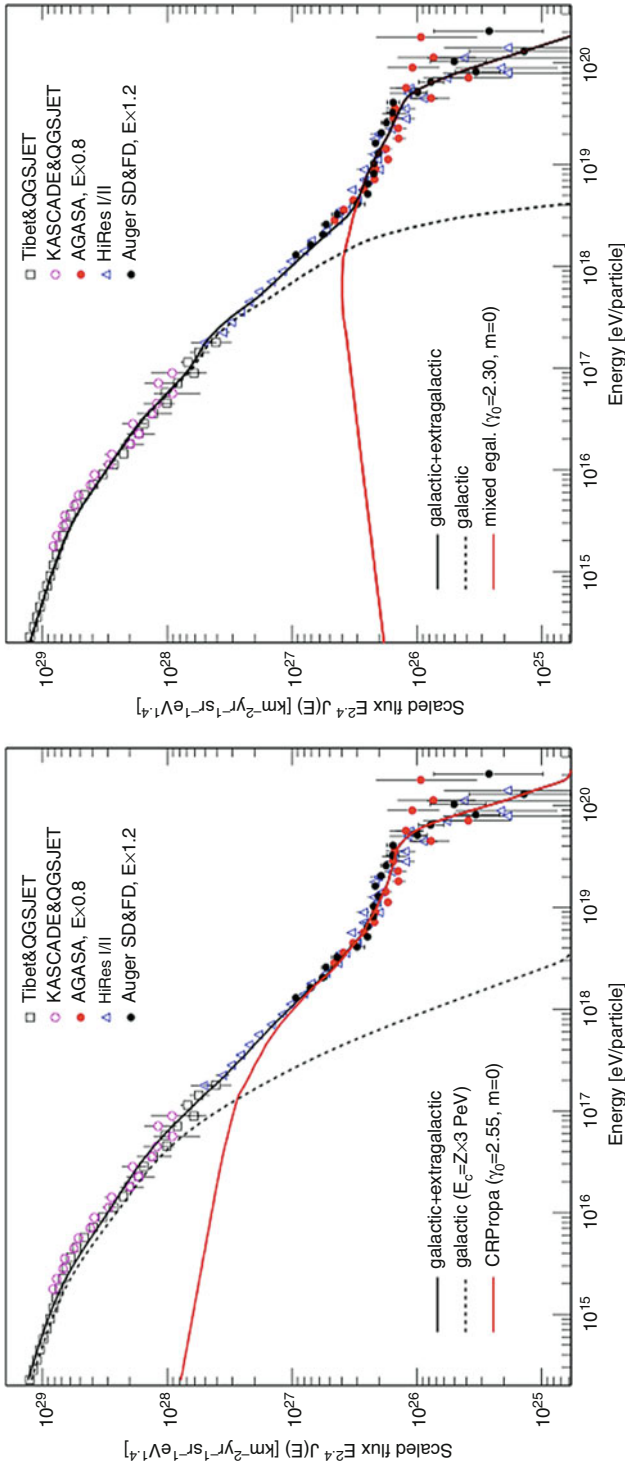
As outlined in the previous section, in the context of the standard model of cosmic ray origin in supernova remnants, subsequent “knees” or steepenings of the spectra are predicted at $E_{\max} \propto Z \times 10^{15}$ eV, reaching $\approx 8 \times 10^{16}$ eV and above for the iron group. Above several 10^{18} eV, the detected cosmic particles must be of extragalactic origin. This assumption also explains the lack of observation of a strong anisotropy that would be expected for charged particles with a large gyroradius at this energy.

The simplest and most natural way of producing a flattening (an ankle) in the cosmic-ray spectrum is that of intersecting the steep galactic spectrum with a flatter extragalactic one. This is the basic idea behind the *ankle model* (Hillas 2005): the transition from galactic to extragalactic cosmic rays appears around 10^{19} eV, at the crossing of the two spectra, producing also the observed “dip,” a concavity in the spectrum at $10^{18} \lesssim E \lesssim 4 \times 10^{19}$ eV. The generation spectrum of the extragalactic component has a slope between 2.2 and 2.5, as predicted by Fermi acceleration both at nonrelativistic and ultrarelativistic shocks, even if these slopes are rather model dependent (Lemoine and Revenu 2006). However, it is clear that additional mechanisms able to accelerate the galactic component beyond the iron knee have to be introduced (Bell and Lucek 2001) in the ankle model, in order to fill the gap between the iron knee and the onset of the extragalactic cosmic rays.

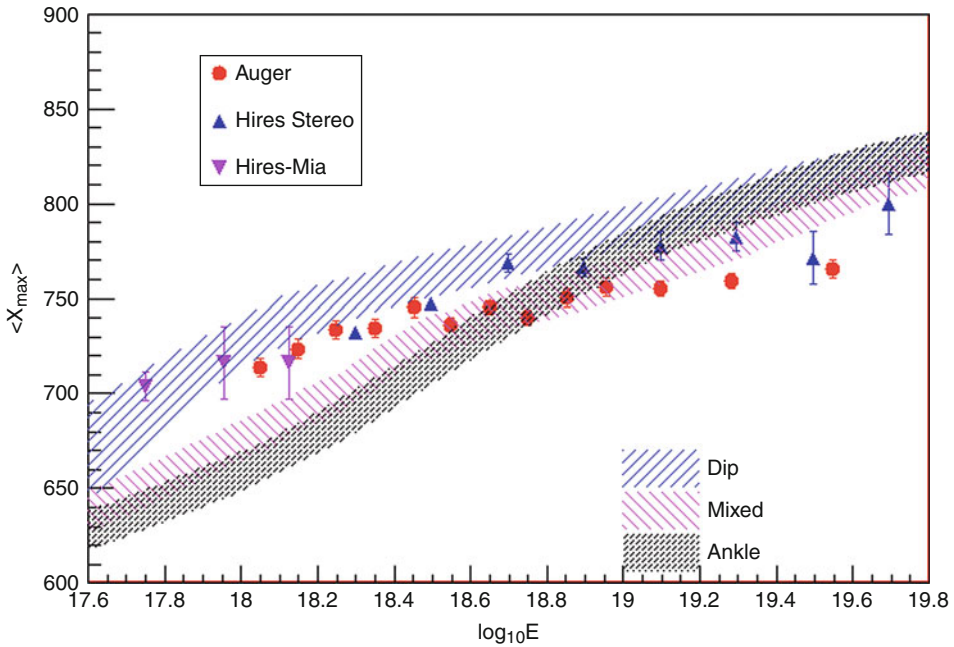
In the *dip model* (Berezinsky et al. 2006), the dip is supposed to be produced by the e^+e^- pair production in the interactions between extragalactic protons (after their escape from the sources (Sigl and Armengaud 2005)) and CMB photons. The transition takes place at the energy at which the adiabatic energy losses due to the expansion of the universe equal the pair production ones, that is, around the so-called second knee ($E \approx 5 \times 10^{17}$ eV); the propagated spectrum and the dip do not depend on the hypotheses on source evolution, although both the beginning of the dip region and the energy at which the transition is completed do. The model requires an almost pure proton composition, with a maximum allowed contamination from He of $\approx 10\%$; at these energies, in fact, the pair production cross section is relevant only for protons, being proportional to the energy/nucleon. A broken power-law generation spectrum is needed to avoid a too large emissivity when extrapolated at lower energy (Berezinsky et al. 2004).

The *mixed composition model* (Allard et al. 2007) assumes that the extragalactic cosmic ray source composition is mixed and similar to that of the galactic cosmic-rays. The observed spectrum can be reproduced by assuming a source spectrum $E^{(-2.2/2.3)}$; the transition region covers energies up to the ankle, while the galactic component extends up to more than 10^{18} eV. Interestingly, the same source spectrum describes low-energy cosmic-ray data, in line with the ideas of holistic models (Parizot 2004), which propose that cosmic rays of any energies be produced by the same sources. A possible difficulty of the mixed model is the fact that the single-element spectra cut off at energies proportional to their mass A . The composition could be dominated by protons below 10^{18} eV, unlike in the ankle model case.

Energy spectrum, composition, and anisotropy are used to discriminate among the different models. The first one is the best measured, but (as shown in  Fig. 14-29 for the dip and mixed composition cases) all models give a good description of the all-particle energy spectrum. In particular, there is an impressive agreement of the dip predictions with the spectrum in the ankle region. A very small level of anisotropy below 1 EeV and isotropy when all cosmic rays become extragalactic can be eventually expected in the transition region.



■ Fig. 14-29 The measured all-particle flux compared to the dip and the mixed composition models described in the text (Unger 2008)

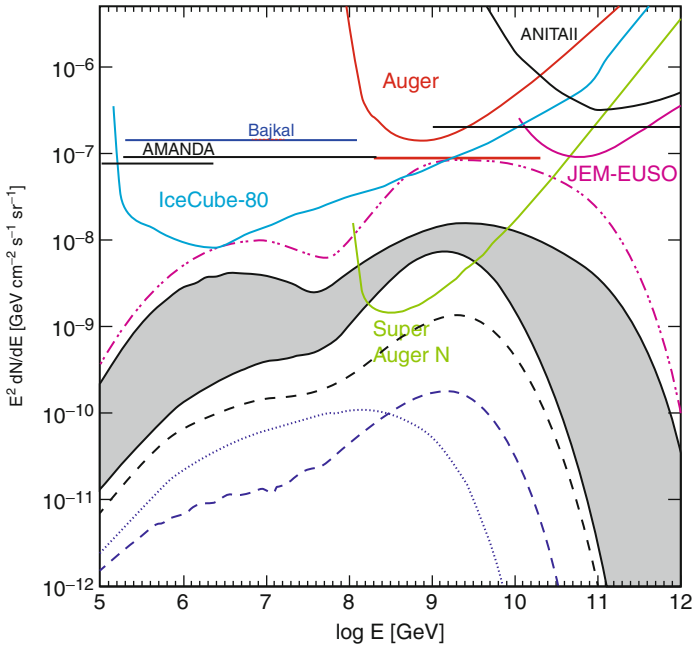


■ Fig. 14-30

$\langle X_{\max} \rangle$ vs. energy compared to model predictions (Adapted from Unger (2008), Auger data from Abraham et al. (2010c))

The study of composition seems to be more efficient in discriminating among the different hypotheses. According to the dip model, the $\langle X_{\max} \rangle$ evolution with energy is steep till the transition ends, becoming then flatter and corresponding to extragalactic protons. In the mixed model, the transition is wider and the evolution of $\langle X_{\max} \rangle$ is less steep, going from the heavy galactic part to the mixed but light extragalactic composition. The experimental data are shown in Fig. 14-30, together with the expectations for proton and iron primaries from two different hadronic interaction models and the predictions described above. None of the three models can satisfactorily reproduce the data; in the case of the mixed composition, the higher number of free parameters allows an easier adjustment.

Another possible key to explore the transition region is that of cosmogenic neutrinos. They originate from the decay of charged pions and neutrons produced in photopion and photonuclear interactions of ultrahigh-energy CRs on the cosmic background photon fields (Berezinsky and Zatsepin 1969). The neutrino fluxes strongly depend on the source density evolution, CR composition, and maximum energy of the accelerated cosmic rays. In Fig. 14-31, they are calculated by Kotera et al. (2010) for different model parameters, distinguishing three possible regions: strong evolution for sources and pure proton composition (pink dot-dashed curve), uniform evolution with pure iron injection and iron-rich composition (blue lines), and intermediate models (gray area). In the same figure, the existing limits from Auger (Abraham et al. 2009) and ANITA-II (Gorham et al. 2010) and the upcoming experimental neutrino sensitivities are also shown. The model prediction spreads a wide range of possible fluxes, with an estimated uncertainty $\approx 50\%$, and can be even further affected by the inclusion of the effects of extragalactic



■ Fig. 14-31

Predicted cosmogenic neutrino flux (Modified from Kotera et al. (2010), see legend therein)

magnetic fields or of galactic strongly magnetized regions. As pointed out in Berezhinsky et al. (2011), at ultrahigh energy some of these scenarios could be excluded by Fermi-LAT results on the diffuse extragalactic gamma ray background (Abdo et al. 2010b); if so, the detection of the UHE neutrinos would require an increase of at least a factor $\simeq 10$ in the current experiments sensitivity. On the other hand, a positive detection of a neutrino flux in the UHE domain, together with PeV measurements, would give information about the galactic to extragalactic transition and the source composition models.

5 The Measure of the Anisotropy

The study of cosmic-ray anisotropies and their evolution with energy is a powerful tool to study propagation properties and sources.

At energies below 10^{17} eV, the cosmic-ray trajectories are bent and isotropized by the galactic magnetic field. Large-scale anisotropies can arise due to the density gradients produced by the propagation of cosmic rays in the galactic magnetic field, described in ▶ Sect. 2, and the high statistics would allow to detect even a small degree of anisotropy. For increasing energy, when the particle gyroradius becomes comparable to the galactic disk thickness, we can expect anisotropy if the sources are distributed in the galactic disk. At the highest energies, no large-scale modulation would be expected, but we could observe events coming from nearby sources (within the GZK horizon) since the magnetic deflections of the particle trajectories are small.

5.1 Large Scale Anisotropy

The simplest anisotropy signal that can be looked for is a dipole in a given direction \vec{j} , which gives rise to an intensity $I(\vec{u}) = I_0 + I_1 \cdot \vec{j}$, with amplitude $(I_{\max} - I_{\min}) / (I_{\max} + I_{\min}) = I_1 / I_0$.

EAS arrays, due to the Earth rotation, operate uniformly with respect to the sidereal time, so that the shower detection and reconstruction are dependent only on declination. Data are generally analyzed in right ascension only, fixing the declination band, due to the difficulty to define the dependence of the detector exposure on declination.

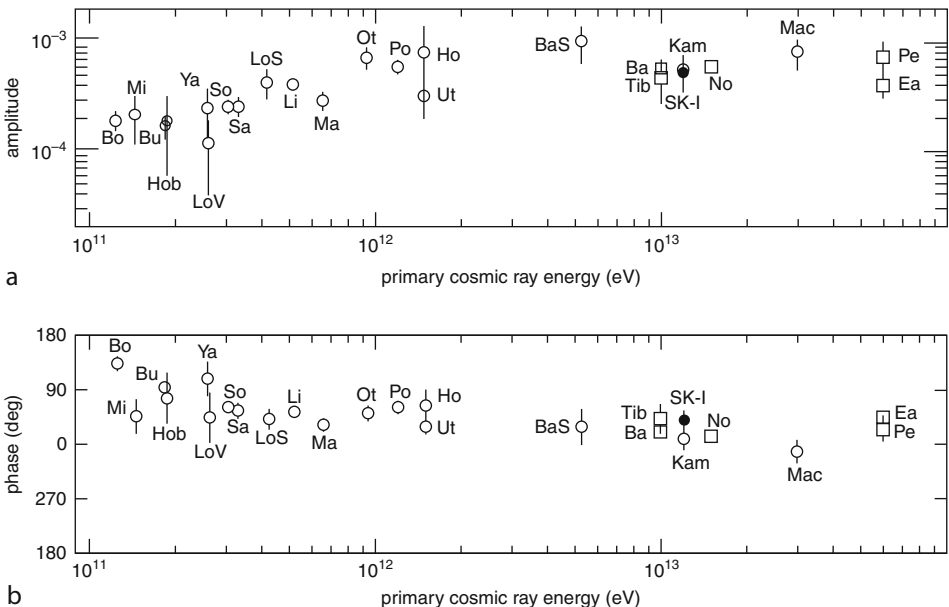
The standard technique to look for large scale anisotropies is the Rayleigh method (Linsley 1975), which, performing an harmonic analysis, allows the amplitude A_k and the phase ϕ_k of the k -th harmonic to be extracted. If α_i, δ_i are the galactic coordinates of the i -th of N events

$$A_k = \sqrt{a_k^2 + b_k^2} \quad \phi_k = \arctan(b_k/a_k), \quad (14.26)$$

where $a_k = \frac{2}{N} \sum_{i=1}^N \cos(k\alpha_i)$ and $b_k = \frac{2}{N} \sum_{i=1}^N \sin(k\alpha_i)$ and the probability of detecting a spurious amplitude by chance is $P(\geq A_k) = \exp(-NA_k^2/4)$.

From the experimental point of view, the measure is feasible only with large collecting areas and long-term observations. The detectors must be uniform in time and area and operate continuously. Systematic effects linked to the atmospheric temperature and pressure variations have to be carefully taken into account.

Anisotropy measurements are obtained either by underground muon observatories (Ambrosio et al. 2003; Guillian et al. 2007), below 10^{13} eV, or by EAS arrays above this energy. A compilation of the results is shown in [Fig. 14-32](#): the measured amplitudes from 10^{11} to 10^{13} eV amount to few times 10^{-4} .




■ Fig. 14-32

Amplitude and phase of anisotropy below 100 TeV (From Guillian et al. (2007), see legend therein)


The EAS-TOP experiment demonstrated that the main features of anisotropy are similar at 10^{14} eV to those measured at lower energies and extended the measurements up to 4×10^{14} eV where $A_{\text{sid}} = (2.6 \pm 0.8) \times 10^{-4}$ and $\phi_{\text{sid}} = (0.4 \pm 1.2) h$ (Aglietta et al. 2009). The analysis was performed adopting a different method, based on the counting rate differences between eastward and westward directions, so removing variations of atmospheric origin.

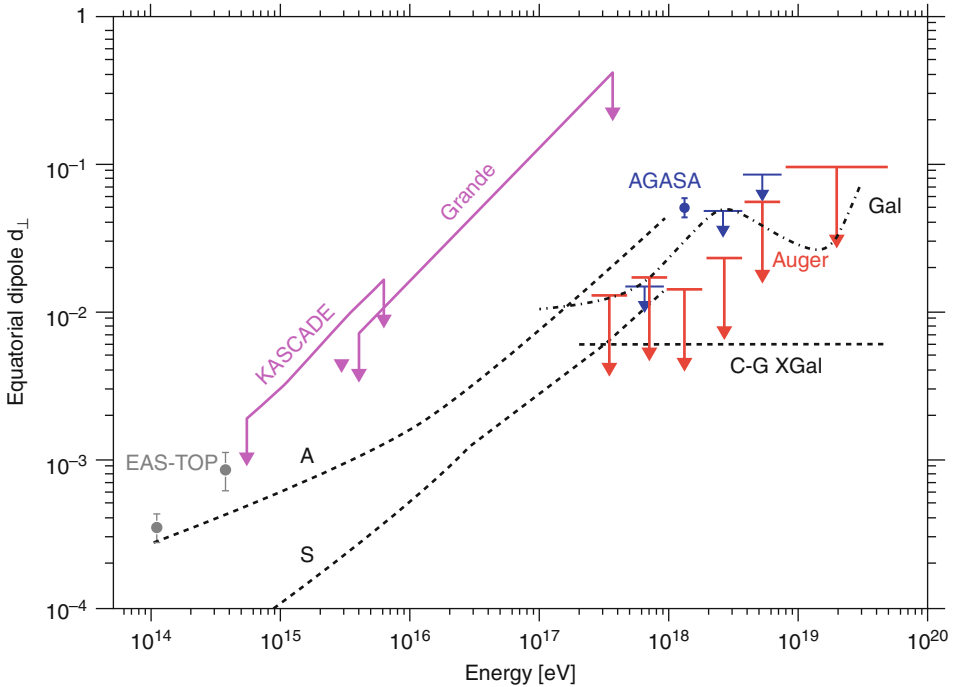
A large scale anisotropy signal is also expected at all energies due to the motion of the observer with velocity \vec{V} with respect to a locally isotropic flux of cosmic rays, the Compton-Getting effect (Compton and Getting 1935). For a CR particle spectrum $I(E) \propto E^{-\gamma}$, the amplitude is $\Delta I/I = \frac{V}{c}(\gamma+2)\cos\theta$, where θ is the angle between the arrival direction of CRs and the moving direction of the observer. For example, the orbital motion of the Earth with velocity $\vec{V} \simeq 30 \text{ km s}^{-1}$ leads to a dipolar anisotropy which can be as large as $\simeq 5 \times 10^{-4}$ (assuming a CR spectrum slope $\gamma = 3$). The expected rate modulation in solar time has in fact been measured at 10 TeV by EAS-TOP (Aglietta et al. 1996) and Tibet (Amenomori et al. 2004).

A similar effect could be expected from the motion of the solar system around the galactic center at $\vec{V} \simeq 220 \text{ km s}^{-1}$, which would produce an anisotropy of few 10^{-3} if the CR plasma were at rest with respect to the Galaxy. The measured sidereal time modulation is an order of magnitude lower, thus demonstrating that the cosmic-ray plasma corotates with the Galaxy (Amenomori et al. 2006).

At higher energy, the anisotropy would be expected to increase if the knee be due to an increasing efficiency of higher and higher energy cosmic rays to escape the Galaxy; the level of anisotropy could then help in choosing among different propagation models. On the other hand, a knee due to the existence of a maximum acceleration energy in the galactic sources would result in a decrease of the anisotropy amplitude with increasing energy, because of the increasing contribution of the extragalactic isotropic cosmic rays component. At these energies, many groups applied the Rayleigh formalism to study the large scale anisotropy, but only upper limits could be derived (Antoni et al. 2004; Over et al. 2007). The current status of the large scale anisotropy measurements in the knee region and above is shown in  Fig. 14-33. The results are plotted taking into account the sky coverage of each apparatus.

As for the higher energy region, different scenarios for the transition bring to different predictions for the anisotropy: if the transition takes place at the ankle energy, the dominant component at 1 EeV is galactic and a modulation at few percent can be expected due to their escape driven by diffusion and drift. If on the contrary extragalactic cosmic rays appear at lower energies, the large scale distribution of isotropic cosmic rays would be influenced by the relative motion of the observer with respect to the frame of the sources. The cosmological Compton-Getting effect, evaluated in the frame in which the CR distribution is isotropic coincides with the CMB rest frame, is shown at about 0.6%. Three times the present statistics would be needed to be sensitive to such amplitude.

The AGASA group reported a large scale anisotropy with dipole-like modulation in right ascension of amplitude $\simeq 4\%$ for a dipole oriented near the galactic center at $E \simeq 1\text{--}2$ EeV, using 11 years of data (Hayashida et al. 1999). The Pierre Auger Observatory (Abreu et al. 2011) studied the sidereal anisotropies using both the Rayleigh and the East-West differential methods, deriving bounds on the first harmonic amplitude at the 1% level at EeV energies. As an example of comparison with models, the expectation derived for the dipole amplitude from diffusion depending on different galactic magnetic field models and source distribution (Candia et al. 2003) is also shown in  Fig. 14-33 (calculated assuming transition at the ankle). The Pierre Auger data are already excluding models where an antisymmetric halo magnetic field is used (indicated with “A”), while nothing can at the moment be said about the “S” model, where the field is symmetric.



■ Fig. 14-33

Upper limit on the anisotropy amplitude as a function of energy (Abreu et al. 2011)

5.2 Point Sources

Clustering at small and intermediate scale can also be searched for in data, to derive information about the number of sources and their distribution. The standard technique is that of comparing the measured arrival direction distribution in equatorial coordinates with an isotropic background distribution, and different methods can be used to derive the signal strength and significance (Mollerach 2009).

Single sources have also been searched for by many groups: no excess of particles from point sources has been detected by KASCADE (Antoni et al. 2004) above 0.3 PeV. In particular, the possible signal from the Monogem ring (a supernova remnant suggested to be the possible single source of galactic cosmic rays (Erlykin and Wolfendale 2004)) has not been confirmed, either by KASCADE or Tibet (Amenomori et al. 2005).

Due to the high density of stars and to the presence of a supermassive black hole, the galactic center is considered a possible site of acceleration for galactic cosmic rays. Reports from AGASA indicate a 4.5σ excess of cosmic rays in the EeV energy range; this result was confirmed by the Fly's Eye and SUGAR detectors (Bellido et al. 2001; Bird et al. 1999), which both claimed for an excess in the direction of the galactic center (the latter with an offset with respect to AGASA). The claim was not supported by the Auger collaboration from the analysis of all data above $10^{18.5}$ eV in the same region of the sky, with statistics much greater than those of previous experiments (Abraham et al. 2007).

The study of discrete sources of CRs, if any, can be best performed exploiting the highest CR energies, since the isotropic background of distant sources would in this case be eliminated by the GZK cutoff and the ultrahigh energy would guarantee sufficient magnetic rigidity for the particles to trace back to their sources. The extremely large exposures and small angular accuracy are the challenge of the ongoing and future experiments aiming to perform this study. Being out of the scope of this review, the present results and discussion pertaining to the anisotropy of extragalactic cosmic rays will not be discussed here and we address the reader to the review of Sommers and Westerhoff (2009).

6 The Future of Cosmic-Ray Astrophysics

The astrophysics of galactic cosmic rays is living a flourishing era, thanks to the wealth of experiments of different conception: ground based, balloon-borne, satellites, or installed on the International Space Station.

Present-day results indicate how powerful and complex is this field in the understanding of the physics and astrophysics of the Milky Way. From one side, we dispose of phenomenological models able to reproduce on a qualitative ground – and in some cases also quantitatively – many different data on cosmic nuclei. On the other side, we still lack a precise modelization of the magnetic Galaxy and of the interactions occurring to CRs in their wandering from their sources to our planet. The understanding of CR data is found on the physics and astrophysics of their sources: mostly SNRs, with some contribution from pulsars. We need a careful – morphological, energetic – study of SNRs in connection with γ -ray data from ground-based as well as space-based telescopes.

By means of direct measurements, both on satellites and balloons, we understood that galactic cosmic rays produced accelerating seeds mainly composed by refractory elements, producing a composition at source very similar to the solar system one, and that $\approx 20\%$ of the sources of galactic cosmic rays are to be found in OB associations.

The diffusion coefficient, which is in principle connected with the inhomogeneities of the galactic magnetic field, can be determined only from the measured spectra of CRs, the best indicator being the boron-to-carbon ratio. Despite the recent experimental achievements, it remains of the utmost importance to dispose of more accurate data on B/C from hundreds of MeV/n to at least the TeV/n region and to increase the statistics in the highest energy region.

Crucial tests of diffusive models are some β -decaying isotopes, the most relevant being $^{10}\text{Be}/^9\text{Be}$. Data are still scarce and very limited in energy. It would be desirable having a measured spectrum extending from around GeV/n up to hundreds (tens at least) GeV/n.

The case of antimatter has received an enormous improvement in the very last years. Antiprotons have been shown to be powerful tests of galactic propagation. Their high agreement with data and the small theoretical uncertainties make this rare species a gauge for astrophysical, nonthermal contributions or exotic contributions from dark matter particles in the galactic halo. It is amazing to notice that in some cases the strongest constraints to particle dark matter models come from the cosmic antiproton data.

The recent measurements of the positron fraction and electron absolute flux clearly indicate that the interpretation of the data requires a careful understanding of the astrophysics of

sources located few kpc around the solar system. Leptons from SNRs and/or pulsars could have registered feeble footprints on the cosmic flux arriving at Earth.

Future direct measurements of cosmic rays will also need to extend the investigation to the high-energy region, towards to knee, in order to establish a benchmark for the cosmic ray composition and so reduce the free parameters in the evaluation of the composition at higher energies.

The requirement of higher statistics, at least ten events per particle type above 10^{15} eV, implies the use of instruments with larger apertures and longer operation time. Part of this goal will be achieved with the new technology at the base of the ultra-long duration balloon flight project, which is currently extensively tested with the aim of producing balloons able to fly for 100 days. However, their limitations in weight and exposure will not allow to reach energies \gtrsim few 10^{14} eV.

In general, the all-particle flux is measured with good agreement among the experiments up to the highest energies, with some evidence of a proportionality of the cutoff energies for each element to the rigidity, even if a dependence on the mass A cannot be excluded. The data show that the proton spectrum is steeper compared to the helium and CNO groups and that there is a dominance of the helium component in the knee region. The results on the composition of single or groups of elements show clearly that the actual limitations are not statistical, but depend mainly on the insufficient knowledge of the characteristics of hadronic interactions, thus underlining the importance of a strict and mutually benefitting link between astroparticle and particle physics. The future results from LHC at CERN will be most helpful in benchmarking the extrapolations to the highest energies.

The investigation of the region above the knee towards the transition to extragalactic cosmic rays requires detectors with large areas but with smaller spacing compared to the arrays studying UHE cosmic rays. As for the lower knee region, the requirement is that of employing complementary techniques, so to detect as many components of showers as possible and cross-check their systematics. Numerous new projects have been designed and are now taking data or are under construction. KASCADE-Grande (Apel et al. 2010), Tunka-133 (Budnev et al. 2007), and Ice-Top (Waldenmaier 2008) aim at exploring the end of the galactic spectrum, eventually detecting the iron knee.

Two enhancements in the Pierre Auger Observatory, the Auger Muon and Infill AMIGA (Platino et al. 2009), and the high-elevation fluorescence telescopes HEAT (Kleifges et al. 2009) will allow to explore the transition region from 10^{17} up to 10^{19} eV. They will give us a powerful tool to clarify the problem of the transition from galactic to extragalactic cosmic rays and complement the measurements at the highest energies.

Intense research and development activity is going on to detect the radio-emission from EAS that could enable both an increase of the statistics and a reduction of systematic uncertainties on the determination of the air shower properties. Progress has been made in recent years by the LOPES and CODALEMA groups (Ardouin et al. 2006; Falcke et al. 2005), while radiodetection at higher energies is under scrutiny in Auger (van den Berg et al. 2009).

Acknowledgments

The authors would like to thank many of their colleagues for the interesting and fruitful discussions. Special thanks are due to A.A. Watson, for the many suggestions.

Cross-References

- [Gamma-Ray Emission of Supernova Remnants and the Origin of Galactic Cosmic Rays](#)
- [Mass Distribution and Rotation Curve in the Galaxy](#)

References

- Aad, G., et al. 2011, *Nat. Commun.*, 2, 463
- Abbasi, R., et al. 2004, *Phys. Rev. Lett.*, 92, 151101
- Abbasi, R., et al. 2005, *ApJ*, 910, 622
- Abbasi, R., et al. 2008, *Phys. Rev. Lett.*, 100, 101101
- Abdo, A. A., et al. 2010a, *Science*, 327, 1103
- Abdo, A. A., et al. 2010b, *Phys. Rev. Lett.*, 104, 101101
- Abdo, A. A., et al. 2011, *ApJ*, 734, 28
- Abe, K., et al. 2008, *Phys. Lett. B.*, 670, 103
- Abe, et al., 2012, *Phys. Rev. Lett.*, 108, 051102
- Abraham, J., et al. 2004, *Nucl. Instrum. Meth. A*, 523, 50
- Abraham, J., et al. 2007, *Astropart. Phys.*, 27, 244
- Abraham, J., et al. 2008, *Phys. Rev. Lett.*, 101, 144
- Abraham, J., et al. 2009, *Phys. Rev. D*, 79, 102001
- Abraham, J., et al. 2010a, *Nucl. Instrum. Meth. A*, 620, 227
- Abraham, J., et al. 2010b, *Phys. Lett. B*, 685, 239
- Abraham, J., et al. 2010c, *Phys. Rev. Lett.*, 104, 091101
- Abreu, P., et al. 2011, *Astropart. Phys.*, 34, 627
- Abu-Zayyad, T., et al. 2001, *ApJ*, 557, 686
- Achterberg, A., et al. 2006, *Astropart. Phys.*, 26, 155
- Ackermann, M., et al. 2010, *Phys. Rev. D*, 82, 092004
- Ackermann, M., et al. 2012, *Phys. Rev. Lett.*, 108, 011103
- Adinolfi-Falcone, R., et al. 1999, *Nucl. Instrum. Meth. A*, 420, 117
- Adriani, O., et al. 2008, *J. Instr.*, 3, 8006
- Adriani, O., et al. 2009a, *Nature*, 458, 607
- Adriani, O., et al. 2009b, *Phys. Rev. Lett.*, 102, 051101
- Adriani, O., et al. 2009c, *Nature*, 458, 607
- Adriani, O., et al. 2010a, *Astropart. Phys.*, 34, 1
- Adriani, O., et al. 2010b, *Phys. Rev. Lett.*, 105, 121101
- Adriani, O., et al. 2011, *Phys. Rev. Lett.*, 106, 201101
- Afanasyev, B. N. 1993, in *Proceedings of the Tokyo Workshop on Techniques for the Study of Extremely High Energy Cosmic Rays*, ed. M. Nagano, (Tokyo: Institute for Cosmic Ray Research, University of Tokyo), 35
- Aglietta, M., et al. 1989, *Nucl. Instrum. Meth. A*, 277, 23
- Aglietta, M., et al. 1991, in *Proceedings of the 23rd International Cosmic Ray Conference*, Dublin, Ireland, 2, 708
- Aglietta, M., et al. 1993, *Nucl. Instrum. Meth. A*, 336, 310
- Aglietta, M., et al. 1996, *ApJ*, 470, 501
- Aglietta, M., et al. 1999, *Astropart. Phys.*, 10, 1
- Aglietta, M., et al. 2004, *Astropart. Phys.*, 21, 223
- Aglietta, M., et al. 2009, *ApJ Lett.*, 692, L130
- Aguilar, M., et al. 2002, *Phys. Rep.*, 366, 331
- Aharonian, F., et al. 2004, *A&A*, 425, L13
- Aharonian, F., et al. 2006, *A&A*, 449, 223
- Aharonian, F., et al. 2007, *Phys. Rev. D*, 75, 042004
- Ahlen, M., et al. 1993, *Nucl. Instrum. Meth. A*, 324, 337
- Ahn, H. S., et al. 2010, *ApJ*, 715, 1400
- Aielli, G., et al. 2009, *Phys. Rev. D*, 80, 92004
- Alcaraz, J., et al. 2000, *Phys. Lett. B*, 484, 10
- Allard, D., et al. 2007, *Astropart. Phys.*, 27, 61
- Amato, E., & Blasi, P. 2006, *MNRAS*, 371, 1251
- Ambrosio, M., et al. 2003, *Phys. Rev. D*, 67, 042002
- Amenomori, M., et al. 2004, *Phys. Rev. Lett.*, 93, 061101
- Amenomori, M., et al. 2005, *ApJ*, 635, L53
- Amenomori, M., et al. 2006, *Science*, 314, 439
- Amenomori, M., et al. 2011, *Astropart. Space Sci. Trans.*, 7, 15
- Anders, E., & Grevesse, N. 1989, *Geoch. Cosmoch. Acta*, 53, 197
- Antchev, G., et al. 2010, *Nucl. Instrum. Meth. A*, 617, 62
- Antoni, T., et al. 2004, *ApJ*, 604, 687
- Antoni, T., et al. 2005, *Astropart. Phys.*, 24, 1
- Apel, W., et al. 2007, *J. Phys. G: Nucl. Part. Phys.*, 34, 2581
- Apel, W., et al. 2009, *Astropart. Phys.*, 31(2), 86
- Apel, W., et al. 2010, *Nucl. Instrum. Meth. A*, 620, 202
- Apel, W., et al. 2011, *Phys. Rev. Lett.*, 107, 171104
- Aramaki, T., et al. 2010, *Adv. Space Res.*, 46, 1349
- Ardouin, D., et al. 2006, *Astropart. Phys.*, 26, 341
- Arqueros, F., et al. 2000, *A&A*, 359, 682
- Arqueros, F., et al. 2008, *Nucl. Instrum. Meth. A*, 597, 23
- Asakimori, K., et al. 1998, *ApJ*, 502, 278
- Asaoka, Y., et al. 2002, *Phys. Rev. Lett.*, 88, 051101
- Askaryan, G. 1962, *Soviet Phys. JETP*, 14, 441
- Atkins, R., et al. 2004, *Astropart. Phys.*, 608, 680
- Auger, P., et al. 1938, *Comptes Rendus*, 206, 1721
- Ave, M., et al. 2003, *Astropart. Phys.*, 19, 47
- Baltrusaitis, R., et al. 1985, *Nucl. Instrum. Meth. A*, 240, 410

- Baltrusaitis, R., et al. 1999, *Nucl. Instrum. Meth. A*, 420, 117
- Barwick, S., et al. 1997a, *Nucl. Instrum. Meth. A*, 400(1), 34
- Barwick, S. W., et al. 1997b, *ApJ*, 482, L191+
- Bell, A., & Lucek, S. 2001, *MNRAS*, 321, 433
- Bell, C. 1974, *J. Phys. A*, 12, 990
- Bellido, J., et al. 2001, *Astropart. Phys.*, 15, 167
- Berezhko, E. G., & Völk, H. J. 2000, *ApJ*, 540, 923
- Berezhko, E. G., Yelshin, V. K., & Ksenofontov, L. T. 1994, *Astropart. Phys.*, 2, 215
- Berezhko, E., et al. 1999, *JETP*, 89, 391
- Berezinskii, V. S., Bulanov, S. V., Dogiel, V. A., & Ptuskin, V. S. 1990, in *Astrophysics of Cosmic Rays*, ed. V. L. Ginzburg (Amsterdam: North-Holland)
- Berezinsky, V., & Zatsepin, G. 1969, *Phys. Lett. B*, 28, 423
- Berezinsky, V., et al. 2004, *Astropart. Phys.*, 21, 617
- Berezinsky, V., et al. 2006, *Phys. Rev. D*, 74, 043005
- Berezinsky, V., et al. 2011, *Phys. Lett. B*, 695, 13
- Bergstrom, L., Edsjo, J., & Ullio, P. 1999, *ApJ*, 526, 215
- Bertaina, M., Navarra, G., Battistoni, G., & Stammer, A. 2008, *J. Phys.: Conf. Ser.*, 120, 062023
- Bertone, G., Hooper, D., & Silk, J. 2005, *Phys. Rep.*, 405, 279
- Biermann, P. 1993, *A&A*, 271, 649
- Bird, D., et al. 1993, *Phys. Rev. Lett.*, 71, 3401
- Bird, D., et al. 1994, *ApJ*, 424, 491
- Bird, D., et al. 1999, *ApJ*, 511, 739
- Blasi, P. 2005, *Mod. Phys. Lett. A*, 20, 3055
- Blasi, P., & Amato, E. 2011, *subm. J. Cosmology Astropart. Phys.*
- Blasi, P., & Serpico, P. D. 2009, *Phys. Rev. Lett.*, 103, 081103
- Blasi, P., Amato, E., & Caprioli, D. 2007, *MNRAS*, 375, 1471
- Blümer, J., et al. 2009, *Progr. Part. Nucl. Phys.*, 63, 293
- Boezio, M., et al. 2000, *ApJ*, 532, 653
- Boezio, M., et al. 2001, *ApJ*, 561, 787
- Bothe, W., & Kolhorster, W. 1929, *Phys. Zeitschr.*, 86, 751
- Bottino, A., Donato, F., Fornengo, N., & Scopel, S. 2004, *Phys. Rev. D*, 70, 015005
- Boyer, J., et al. 2002, *Nucl. Instrum. Meth. A*, 482, 457
- Budnev, N., et al. 2007, in *Proceedings of the 30th International Cosmic Ray Conference*, Merida. [arXiv:0801.3037](https://arxiv.org/abs/0801.3037)
- Candia, J., et al. 2002, *Astropart. Phys.*, 17, 23
- Candia, J., et al. 2003, *J. Cosmol. Astropart. Phys.*, 05, 003
- Casse, M., & Goret, P. 1978, *ApJ*, 221, 703
- Chang, J., et al. 2008, *Nature*, 456, 362
- Chardonnet, P., Orloff, J., & Salati, P. 1997, *Phys. Lett. B*, 409, 313
- Chernov, D., et al. 2005, *Int. J. Mod. Phys. A*, 20, 6799
- Chiba, N., et al. 1992, *Nucl. Instrum. Meth. A*, 311, 338
- Clay, J. 1930, *Amst. Proc.*, 33, 711
- Compton, A., & Getting, I. 1935, *Phys. Rev.*, 47, 817
- Dar, A., & De Rujula, A. 2004, *Phys. Rep.*, 405, 203
- Dar, A., & De Rujula, A. 2008, *Phys. Rep.*, 466, 179
- Delahaye, T., Lineros, R., Donato, F., Fornengo, N., Lavalle, J., Salati, P., & Taillet, R. 2009, *A&A*, 501, 821
- Delahaye, T., Lavalle, J., Lineros, R., Donato, F., & Fornengo, N. 2010, *A&A*, 524, A51+
- Derbina, V., et al. 2005, *ApJ*, 628, 4
- Donato, F., Fornengo, N., & Salati, P. 2000, *Phys. Rev.*, D62, 043003
- Donato, F., et al. 2001, *ApJ*, 563, 172
- Donato, F., Maurin, D., & Taillet, R. 2002, *A&A*, 381, 539
- Donato, F., Fornengo, N., Maurin, D., & Salati, P. 2004, *Phys. Rev.*, D69, 063501
- Donato, F., Fornengo, N., & Maurin, D. 2008, *Phys. Rev.*, D78, 043506
- Donato, F., Maurin, D., Brun, P., Delahaye, T., & Salati, P. 2009, *Phys. Rev. Lett.*, 102, 071301
- Dova, M. T., et al. 2001. [arXiv:astro-ph/0112191](https://arxiv.org/abs/astro-ph/0112191)
- Drury, L. O., Axford, W. I., & Summers, D. 1982, *MNRAS*, 198, 833
- Drury, L. O., Aharonian, F. A., & Voelk, H. J. 1994, *A&A*, 287, 959
- Drury, L. O., et al. 2001. [arXiv:astro-ph/0106046](https://arxiv.org/abs/astro-ph/0106046)
- Duperray, R., et al. 2005, *Phys. Rev.*, D71, 083013
- Edge, D., et al. 1973, *J. Phys. A*, 6, 1612
- Egorova, V., et al. 2001, *J. Phys. Soc. Jpn.* 70(2001) Suppl.B, 70, 9
- Engelmann, J., et al. 1990, *A&A*, 233, 96
- Engler, J., et al. 1999, *Nucl. Instrum. Meth. A*, 427, 528
- Enomoto, R., et al. 2002, *Nature*, 426, 823
- Erber, T. 1966, *Rev. Mod. Phys.*, 38, 626
- Erlykin, A., & Wolfendale, A. 2001, *J. Phys. G: Nucl. Part. Phys.*, 27, 1005
- Erlykin, A., & Wolfendale, A. 2004, *Astropart. Phys.*, 22, 47
- Erlykin, A., & Wolfendale, A. 2009, in *Proceedings of the 31st International Cosmic Ray Conference*, Lodz (Poland), 0301
- Erlykin, A., & Wolfendale, A. 2011, *Astrophys. Space Sci. Trans.*, 7, 145
- Evoli, C., Gaggero, D., Grasso, D., & Maccione, L. 2008, *J. Cosmology Astropart. Phys.*, 10, 18
- Falcke, H., et al. 2005, *Nature*, 435, 313
- Fermi, E. 1949, *Phys. Rev.*, 75, 1169
- Fowler, J. 2001, *Astropart. Phys.*, 15, 49
- Fowler, J., et al. 2001, *Astropart. Phys.*, 15, 49
- Fuke, H., et al. 2005, *Phys. Rev. Lett.*, 95, 081101
- Fukui, Y., et al. 2003, *Publ. Astron. Soc. Jpn.*, 55, L61

- Gaisser, T. K. 1990, *Cosmic Rays and Particle Physics*, (Cambridge and New York: Cambridge University Press), 292
- Gaisser, T., & Hillas, A. 1977, in *Proceedings of the 15th International Cosmic Ray Conference*, Plovdiv, Bulgaria, 358
- Ginzburg, V., & Syrovatskii, S. 1964, (London/New York: Pergamon Press)
- Ginzburg, V. L., Khazan, I. M., & Ptuskin, V. S. 1980, *Ap&SS*, 68, 295
- Glasmacher, M., et al. 1999, *Astropart. Phys.*, 12, 1
- Globus, N., Allard, D., & Parizot, E. 2008, *A&A*, 479, 97
- Gorham, P. W., et al. 2010, *Phys. Rev. D*, 82, 022004
- Greisen, K. 1960, *Ann. Rev. Nucl. Sci.*, 10, 63
- Greisen, K. 1966, *Phys. Rev. Lett.*, 16, 748
- Guillian, G., et al. 2007, *Phys. Rev. D*, 75, 062003
- Gupta, S., et al. 2009, *Nucl. Phys. B (Proc. Suppl.)*, 196, 153
- Hannestad, S. 1993, *New J. Phys.*, 1, 229
- Haungs, A., et al. 2003, *Rep. Progr. Phys.*, 66, 62
- Hayashida, N., et al. 1999, *Astropart. Phys.*, 10, 303
- Heck, D., et al. 1998, *FZKA Report Forschungszentrum Karlsruhe*, 6019
- Hess, V. 1912, *Phys. Z*, 13, 1084
- Higdon, J., & Lingenfelter, R. 1999, *ApJ*, 523, L61
- Higdon, J., et al. 1998, *ApJ*, 509, L33
- Hillas, A. 1982, *J. Phys. G*, 8, 1475
- Hillas, A. 1984, *Ann. Rev. Astron. Astrophys.*, 22, 425
- Hillas, A. 2005, *J. Phys. G*, J31, R95
- Hillas, A., et al. 1971, in *Proceedings of the 12th International Cosmic Ray Conference*, Hobart, 3, 1001
- Hoerandel, J. 2004, *Astropart. Phys.*, 21, 241
- Hoerandel, J., et al. 2009, in *Proceedings of the 31st International Cosmic Ray Conference*, Lodz, 0227
- Ipavich, F. M. 1975, *ApJ*, 196, 107
- Ivanov, A. 2010, *ApJ*, 712, 746
- Jelley, J., et al. 1965, *Nature*, 205, 327
- Johnson, H. E., & Axford, W. I. 1971, *ApJ*, 165, 381
- Jones, F. C., Lukasiak, A., Ptuskin, V., & Webber, W. 2001, *ApJ*, 547, 264
- Kamata, K., & Nishimura, J. 1958, *Prog. Theo. Phys.*, 6, 93
- Karakula, S., & Tkaczyk, W. 2004, *Astropart. Phys.*, 6, 108
- Karle, A., et al. 1995, *Astropart. Phys.*, 3, 321
- Kazanas, D., & Nikolaidis, A. 2001, *Gen. Rel. Grav.*, 35, 1117
- Khristiansen, G., et al. 1977, in *Proceedings of the 15th International Cosmic Ray Conference*, Plovdiv, 8, 148
- Klages, H., et al. 1997, *Nucl. Phys. B (Proc. Suppl.)*, 52, 92
- Kleifges, M., et al. 2009, in *Proceedings of the 31st International Cosmic Ray Conference*, Lodz, 410
- Kobayakawa, K., et al. 2002, *Phys. Rev. D*, 66, 083004
- Kolhorster, W. 1925, *Phys. Z*, 26, 654
- Kolhorster, W. 1938, *Nature*, 26, 576
- Kopenkin, V., & Sinzi, T. 2009, *Phys. Rev. D*, 79, 72011
- Kotera, K., & Olinto, A. 2011, *Ann. Rev. Astron. Astrophys.*, 49, 119
- Kotera, K., Allard, D., & Olinto, A. V. 2010, *J. Cosmology Astropart. Phys.*, 10, 13
- Kulikov, G., & Khristiansen, G. 1958, *JETP*, 35, 635
- Lagage, P. O., & Cesarsky, C. J. 1983a, *A&A*, 118, 223
- Lagage, P. O., & Cesarsky, C. J. 1983b, *A&A*, 125, 249
- Lagutin, A., et al. 2001, *Nucl. Phys. B, Proc. Suppl.*, 97, 267
- Larsen, C., et al. 2001, in *Proceedings of the 27th International Cosmic Ray Conference*, Hamburg, 134
- Lawrence, M. A., et al. 1991, *J. Phys. G*, 17, 733
- Lemoine, M., & Revenu, B. 2006, *MNRAS*, 366, 635
- Link, J., et al. 2009, *ApJ*, 697, 2083
- Linsley, J. 1963, *Phys. Rev. Lett.*, 10, 146
- Linsley, J. 1975, *Phys. Rev. Lett.*, 34, 1530
- Lodders, K. 2003, *ApJ*, 591, 1220
- Lorenz, E. 1996, *Space Sci. Rev.*, 75, 169
- Maeno, T., et al. 2001, *Astropart. Phys.*, 16, 121
- Malkov, M. A., & Drury, L. O. 2001, *Rep. Prog. Phys.*, 64, 429
- Mannheim, K., & Schlickeiser, R. 1994, *A&A*, 286, 983
- Matthews, J. 2005, *Astropart. Phys.*, 22, 387
- Maurin, D., Donato, F., Taillet, R., & Salati, P. 2001, *ApJ*, 555, 585
- Maurin, D., Taillet, R., & Donato, F. 2002a, *A&A*, 394, 1039
- Maurin, D., et al. 2002b, [arXiv:astro-ph/0212111](https://arxiv.org/abs/astro-ph/0212111)
- Mewaldt, R. A., et al. 2010, *ApJ*, 723, L1
- Meyer, J., Drury, L. O., & Ellison, D. C. 1997a, *ApJ*, 487, 182
- Meyer, J.-P., et al. 1997b, *ApJ*, 487, 182
- Mitchell, J. W., et al. 2005, *Adv. Space Res.*, 35, 135
- Mollerach, S. 2009, *AIP Conf. Proc.*, 1123, 115
- Moskalenko, I. V., & Strong, A. W. 1998, *ApJ*, 493, 694
- Müller, D., et al. 2009, in *Proceedings of the 31st International Cosmic Ray Conference*, Lodz, 914
- Nagano, M. 2009, *New J. Phys.*, 11, 065012
- Nagano, M., et al. 1992a, *J. Phys. G*, 8, 423
- Nagano, M., et al. 1992b, *J. Phys. G*, 18, 423
- Nagano, M., et al. 2004, *Astropart. Phys.*, 22, 235
- Nakamura, K., et al. 2010, *J. Phys. G*, 37, 075021
- Obermeier, A., et al. 2011, *Astrophys. J.*, 742, 14
- Ogio, S., et al. 2003, in *Proceedings of the 28th International Cosmic Ray Conference*, Tsukuba, 1, 315
- Orito, S., et al. 2000, *Phys. Rev. Lett.*, 84, 1078

- Over, S., et al. 2007, in Proceedings of the 30th International Cosmic Ray Conference, Merida
- Pacini, D. 1912, *Il Nuovo Cimento.*, 6(III), 93
- Parizot, E. 2004, in “New Views on the Universe”, Proceedings of the Vth Rencontres du Vietnam, Hanoi. [astro-ph/0501274](https://arxiv.org/abs/astro-ph/0501274)
- Parker, E. N. 1965, *ApJ*, 142, 584
- Petrukhin, A. 2003, *Phys. Atom. Nuclei*, 66, 517
- Platino, M., et al. 2009, in Proceedings of the 31st International Cosmic Ray Conference, Lodz, 184
- Ptitsyna, K., & Troitsky, S. 2010, *Physics Uspekhi*, 53, 691
- Ptuskin, V., et al. 1993a, *A&A*, 268, 726
- Ptuskin, V. S., Rogovaya, S. I., Zirakashvili, V. N., Chuvilgin, L. G., Khristiansen, G. B., Klepach, E. G., & Kulikov, G. V. 1993b, *A&A*, 268, 726
- Putze, A., Derome, L., & Maurin, D. 2010, *A&A*, 516, A66
- Rauch, B., et al. 2009, *ApJ*, 697, 2083
- Rossi, B., & Greisen, K. 1941, *Rev. Mod. Phys.*, 13, 240
- Roulet, E., et al. 2004, *Int. J. Mod. Phys. A*, 19, 1133
- Salati, P., Donato, F., & Fornengo, N. 2010, in Particle Dark Matter: Observations, Models and Searches (Cambridge, UK: Cambridge University Press)
- Schlickeiser, R., & Lerche, I. 1985, *A&A*, 151, 151
- Seo, E. S., & Ptuskin, V. S. 1994, *ApJ*, 431, 705
- Seo, E. S., et al. 2004, *Adv. Spa. Res.*, 33, 1777
- Sigl, G., & Armengaud, E. 2005, *JCAP*, 510, 16
- Sommers, P., & Westerhoff, S. 2009, *New J. Phys.*, 11, 5003
- Stanev, T., et al. 1982, *Phys. Rev. D*, 25, 1291
- Stanev, T., et al. 1993, *A&A*, 274, 902
- Stone, E., et al. 1998, *Spa. Sci. Rev.*, 86, 285
- Strong, A. W., & Moskalenko, I. V. 1998, *ApJ*, 509, 212
- Strong, A. W., Moskalenko, I. V., & Reimer, O. 2004, *ApJ*, 613, 962
- Strong, A. W., Moskalenko, I. V., & Ptuskin, V. S. 2007, *Ann. Rev. Nucl. Part. Sci.*, 57, 285
- Sveshnikova, L., et al. 2003, *A&A*, 409, 799
- Swordy, S. 1995, in Proceedings of the 24th International Cosmic Ray Conference, Roma, 2, 697
- Takahashi, Y., et al. 2009, *New J. Phys.*, 11, 065009
- Takeda, M., et al. 2003a, *Astropart. Phys.*, 19, 447
- Takeda, M., et al. 2003b, *Astropart. Phys.*, 19, 447
- Trotta, R., et al. 2011, *ApJ*, 729, 106
- Uchiyama, Y., et al. 2003, *A&A*, 400, 567
- Ulrich, R., et al. 2009, *New J. Phys.*, 11, 065018
- Ulrich, R., et al. 2011, in Proceedings of the 32nd International Cosmic Ray Conference, Beijing
- Unger, M. 2008. [arXiv:0812.2763](https://arxiv.org/abs/0812.2763)
- van den Berg, A., et al. 2009, in Proceedings of the 31st International Cosmic Ray Conference, Lodz, 0232
- Waldenmaier, T. 2008, *Nucl. Instrum. Meth. A*, 588, 130
- Wang, J., et al. 2002, *ApJ*, 564, 244
- Wang, B., et al. 2010, *Sci. China Phys. Mech. Astron.*, 53, 842
- Webber, W. R., Kish, J. C., & Schrier, D. A. 1990, *Phys. Rev.*, C41, 566
- Webber, W. R., Lee, M. A., & Gupta, M. 1992, *ApJ*, 390, 96
- Wiedenbeck, M., et al. 2003, in Proceedings of the 28th International Cosmic Ray Conference, Tsukuba, 4, 1899
- Wiedenbeck, M., et al. 2007, *Space Sci. Rev.*, 130, 415
- Wigmans, R. 2003, *Astropart. Phys.*, 19, 379
- Wischnewski, R., et al. 2005, *Int. J. Mod. Phys. A*, 20, 6392
- Yoshida, T., et al. 2004, *Adv. Space Res.*, 33, 1755
- Zank, G. P. 1989, *A&A*, 225, 37
- Zatsepin, G., & Kuzmin, V. 1966, *J. Exp. Theor. Phys. Lett.*, 4, 78
- Zirakashvili, V. N., & Aharonian, F. A. 2010, *ApJ*, 708, 965
- Zuccon, P., et al. 2009, in Proceedings of the 31st International Cosmic Ray Conference, Lodz, 1273

15 **Gamma-Ray Emission of Supernova Remnants and the Origin of Galactic Cosmic Rays**

F. A. Aharonian

Dublin Institute for Advanced Studies, Dublin, Ireland

Max-Planck-Institut für Kernphysik, Heidelberg, Germany

1	<i>Introduction</i>	790
2	<i>Gamma-Ray Detectors</i>	792
3	<i>Supernova Remnants and Galactic Cosmic Rays</i>	794
4	<i>SN 1006: A Classical Shell-Type Supernova Remnant</i>	796
5	<i>RX J1713.7-3946: A Unique SNR</i>	799
5.1	Challenges of Hadronic Models	804
5.2	Challenges of Leptonic Models	807
6	<i>DSA with Forward and Reversed Shocks Applied to RX J1713.7-3946</i>	808
7	<i>Cas A</i>	814
8	<i>Searching for Galactic Proton PeVatrons</i>	819
9	<i>Gamma-Ray “Echos” from Nearby Molecular Clouds</i>	822
10	<i>Summary</i>	825
	<i>References</i>	825

Abstract: The recent surveys of the Milky Way with space and ground-based gamma-ray detectors revealed hundreds of high energy (HE) and tens of very high energy (VHE) gamma-ray emitters representing several galactic source populations – supernova remnants, giant molecular clouds, star forming regions, pulsars, pulsar wind nebulae, binary systems. The major fraction of these objects remains however unidentified. In this chapter I discuss the astrophysical implications of VHE gamma-ray observations of supernova remnants (SNRs) in the context of the origin of galactic cosmic rays. These observations confirm the earlier theoretical predictions of effective acceleration of multi-TeV particles in young SNRs by strong shock waves. The interpretation of VHE gamma-ray data from several prominent representatives of young SNRs within the so-called hadronic models requires hard energy spectra of protons extending to 100 TeV, with total energy released in relativistic protons and nuclei as large as 10^{50} erg. Formally, this can be considered as an observational proof of the so-called SNR paradigm of the origin of galactic cosmic rays. However, the hadronic models are not free of problems related to interpretation of multi-wavelength properties of these objects. Moreover, in most of the cases the gamma-ray data can be explained by the inverse Compton scattering of electrons which are responsible also for the synchrotron X-radiation of young SNRs. These circumstances prevent us from a firm statement about the contribution of SNRs to the overall flux of galactic cosmic rays. Further observations of young SNRs, especially in the highest energy band (well above 10 TeV), can be crucial in this regard. Quite important are also the complementary observations from massive molecular clouds located within the close proximity of mid-age SNRs.

1 Introduction

Cosmic gamma-rays carry key information about high-energy phenomena in a variety of astrophysical environments. Being a part of modern astrophysics, gamma-ray astronomy is a discipline in its own right. It addresses an impressively broad range of topics related to the nonthermal Universe – acceleration, propagation, and radiation of relativistic particles on different astronomical scales: from compact objects like pulsars (neutron-stars) and microquasars (accreting black holes) to giant radiogalaxies and galaxy clusters (see, e.g., Aharonian (2004)). The energy range of gamma-ray observations reported from galactic sources extends from 100 keV to 100 TeV. While the lower bound characterizes the region of nuclear gamma-ray lines, the upper bound (100 TeV) associates with the highest energy particles accelerated in galactic objects, in particular in young supernova remnants (SNRs) and pulsar wind nebulae (PWN).

The low or MeV energy interval is uniquely linked to several astrophysical phenomena, in particular to nucleosynthesis of heavy elements related to the type Ia Supernovae, Gamma-Ray Bursts, Solar flares, interactions of subrelativistic cosmic rays in the interstellar medium, production and annihilation of positrons, etc. Unfortunately, the sky in low-energy gamma-rays remains an essentially unexplored frontier. The challenge here is connected to the design and construction of sensitive gamma-ray detectors which would provide breakthrough in the field. The combination of several principal factors – the low detection efficiency, the modest angular resolution, and the “heavy” background – severely limit the potential of detectors operating in this energy band. The minimum detectable energy fluxes, even after significant improvements foreseen for the next-generation of low-energy gamma-ray detectors, will remain relatively

modest, hardly better than 10^{-12} erg cm $^{-2}$ s $^{-1}$. Even so, low-energy gamma-rays contain crucial astronomical information that cannot be obtained by other means. This concerns, e.g., the probes of the flux of subrelativistic ($E \leq 100$ MeV) cosmic rays in the Interstellar Medium (ISM) through their prompt deexcitation gamma-ray lines. Such measurements may provide the only direct (model-independent) estimates of the ionization and heating rate of ISM by low-energy cosmic rays (see, e.g., Kozlovsky et al. (2002)). Another important implication of nuclear gamma-ray line emission is related to the measurements of temperature of the nucleonic component of thermal very hot plasmas with $kT_i \geq 1$ MeV formed in strong shock waves or in two-temperature accretion flows close to the solar mass black holes (Aharonian and Sunyaev 1984). And of course, one of the major aspects of MeV gamma-ray astronomy is related to the exploration of the 0.511 MeV line emission due to annihilation of positrons which are copiously produced in various astrophysical scenarios (see, e.g., Prantzos et al. (2011), Churazov et al. (2011)). Therefore any further improvement of the detector performance in this energy band is anticipated to result in confirmation or rejection of earlier theoretical predictions and, hopefully, also in exciting serendipitous discoveries.

Cosmic gamma-ray emission is better explored in the *high* or GeV ($0.1 \text{ GeV} \leq E_\gamma \leq 0.1 \text{ TeV}$) and *very high* or TeV ($0.1 \text{ TeV} \leq E_\gamma \leq 10 \text{ TeV}$) energy intervals. This is explained, among other reasons, by the adequate detection performance in these two energy bands, and by high (in some cases, extremely high) efficiency of acceleration and radiation of particles in diverse variety of astrophysical settings. While observations in the first energy band are obtained with satellite-borne detectors, the second energy interval is explored by ground-based instruments.

Before the launch of the *Fermi Gamma-ray Space Telescope* in May 2008, the high, energy space-based gamma-ray astronomy has been dominated by the results obtained with the *EGRET* telescope aboard the Compton Gamma Ray Observatory. Because of rather modest angular resolution (of order of a few degrees), only two source populations – the Active Galactic Nuclei (AGN) and pulsars – have been clearly identified by *EGRET* as high-energy gamma-ray emitters. The observations with *Fermi* with significantly improved (compared to *EGRET*) flux sensitivity and angular resolution led to the detection of tens of new gamma-ray pulsars and hundreds of gamma-ray emitting AGN, as well as to the discovery of new type galactic and extragalactic gamma-ray sources. Finally, *Fermi* extended the observations of the two components of the diffuse gamma-ray emission related to the Galactic Disk and the isotropic (extragalactic) diffuse gamma-ray background, to energies up to 100 GeV. A number of interesting results, including the detection of synchrotron flares of the Crab Nebula and detection of several variable gamma-ray sources like Cygnus X-3, have been obtained also with the Italian gamma-ray satellite *AGILE*, another new generation high-energy gamma-ray detector with a better performance compared to *EGRET* (Tavani 2010).

One of the most remarkable achievements of recent years in astrophysics was the sudden emergence of very high-energy gamma-ray astronomy as a truly astronomical discipline. The observations conducted by the *HESS*, *MAGIC*, *VERITAS*, and *MILAGRO* groups resulted in the discovery of VHE gamma-ray sources with a number in excess of 100. Remarkably, these sources represent almost all major nonthermal astrophysical source populations, including shell-type Supernova Remnants, Pulsar Wind Nebulae, Star Forming Regions, Giant Molecular Clouds, X-ray Binary Systems, Blazars, Radiogalaxies, and Starburst Galaxies (for a review see, e.g., Aharonian et al. (2008a), Hinton and Hofman (2009)). For many, this success was a big surprise, especially given the rather controversial history of the field over the last four decades (Aharonian 2004). In this regard, a question naturally arises concerning the reasons which made

possible this success. A likely answer to this question perhaps can be formulated as a lucky combination of two independent factors: (1) the practical realization of the great potential of stereoscopic arrays of Imaging Atmospheric Cherenkov Telescopes as effective multifunctional tools for study of cosmic gamma-radiation from ground, and (2) the existence of a large variety of perfectly designed cosmic accelerators – TeVatrons, PeVatrons, and perhaps EeVatrons – the factories of nonthermal relativistic matter where particle acceleration proceeds with efficiency close to the theoretical limits determined by the classical electrodynamics and plasma physics. The recent VHE observations tell us that often the effective particle acceleration is accompanied with creation of favorable conditions for gamma-ray production.

2 Gamma-Ray Detectors

The Fermi Large Area Telescope *Fermi LAT*, presently the most sensitive space-based gamma-ray instrument, is perfectly designed for deep surveys of the sky with an effective field of view of order of 2 sr. Formally, it is claimed to cover four energy decades, from 20 MeV to more than 300 GeV (Michelson et al. 2010). However, because of dramatic reduction of the detection area and worsening of the angular resolution at low energies, the effective *threshold* is rather close to 100 MeV. The angular resolution of *Fermi LAT* below 1 GeV is quite modest, larger than 1° , but it is getting significantly better with energy approaching to 0.1° above 10 GeV. Generally, at such high energies the gamma-ray fluxes are quite low, and the detection area of about $1 \leq m^2$ cannot provide adequate photon statistics, in particular for spectroscopic and morphological studies. With some exceptions, this limits the detection of tiny fluxes of cosmic gamma-rays to energies ≤ 100 GeV. The best performance of the instrument with a sensitivity as good as 10^{-12} erg cm $^{-2}$ s $^{-1}$, angular resolution better than 0.3° and energy resolution better than 10% is achieved at multi-GeV energies.

Space-based instruments cannot offer, at least in the foreseeable future, detection areas exceeding significantly $1 m^2$; this limits dramatically the potential of space-based gamma-ray astronomy for studies of the VHE domain. Fortunately, at such higher energies an alternative method can be used. The method is based on registration of atmospheric showers (initiated by interactions of gamma-rays) either directly or through their Cherenkov radiation. The faint and brief Cherenkov signal which lasts only several nanoseconds can be detected by large optical reflectors equipped with fast multipixel cameras. With a telescope consisting of an optical reflector of diameter $D \approx 10$ m, and a multichannel camera with pixel size $0.1\text{--}0.2^\circ$ and field-of-view $\Theta \geq 3^\circ$, primary gamma-rays of energy ≥ 100 GeV can be collected from distances as large as 100 m. This provides huge detection areas, $A \geq 3 \times 10^4 m^2$, which largely compensate the weak gamma-ray fluxes at these energies. The total number of photons in the registered Cherenkov light image is a measure of energy, the orientation of the image correlates with the arrival direction of the gamma-ray, and the shape of the image contains information about the origin of the primary particle (a proton or photon). The stereoscopic observations of air showers with two or more telescopes located at distances of about 100 m from each other provide quite effective rejection of hadronic showers (by a factor of 100), as well as very good angular (better than 0.1°) and energy (better than 15%) resolutions. At energies around 1 TeV, this results in a minimum detectable energy flux as low as 3×10^{-13} erg cm $^{-2}$ (Aharonian 2004). This is a quite impressive sensitivity even in the standards of advanced branches of astrophysics. For example,

for extended sources with angular size larger than 1 arcmin, the IACT arrays provide a sensitive probe of high-energy gamma-rays, competitive with the potential of the *XMM-Newton* and *Chandra* telescopes in the X-ray band. The sensitivity of IACT arrays is much better than in any other gamma-ray domain, including the GeV energy band, where the sensitivity of *Fermi*, even after dramatic improvement compared to the performance of the previous gamma-ray spaceborne instruments, still cannot compete with the performance already achieved in the TeV energy band. Moreover, thanks to very large collection area, the IACT technique provides large gamma-ray photon statistics even from relatively modest TeV gamma-ray emitters. Coupled with good energy and angular resolutions, the rich photon statistics allows deep morphological, spectral, and temporal studies. This makes the IACT arrays perfect multifunctional and multipurpose astronomical tools for exploration of a broad range of nonthermal objects and phenomena.

The IACT arrays are designed for observations of point-like or moderately extended (with angular size 1° or less) objects with known celestial coordinates. However, the high sensitivity and relatively large ($\geq 4^\circ$) field of view of IACT arrays allow effective all-sky surveys. In particular, a large number of the *HESS* galactic sources (including the ones not yet unidentified) with energy fluxes as low as 10^{-12} erg cm $^{-2}$ s $^{-1}$ have been discovered in the *HESS* survey of the galactic plane. On the other hand, the potential of IACT arrays is relatively limited for the search for very extended structures, like diffuse emission, associated, e.g., with the inner Galaxy. The capability of IACT arrays is limited also for search for solitary events or prompt reaction to the increase of activity of highly variable sources. In this regard, the ground-based detection technique based on direct registration of particles that comprise the extensive air showers (EAS) is a complementary approach to the IACT technique. The traditional EAS technique, based on particle detectors, e.g., scintillators, spread over large areas, works effectively for detection of cosmic rays at ultrahigh energies, $E \geq 100$ TeV. In order to make this technique more relevant to the purposes of gamma-ray astronomy, the detection energy threshold should be reduced by two orders of magnitude. This can be achieved using dense particle arrays or large water Cherenkov detectors located on very high altitudes. The feasibility of both approaches recently has been successfully demonstrated by the Tibet AS $_{\gamma}$ and Argo collaborations, and especially by the Milagro group which has reported statistically significant detections of diffuse multi-TeV gamma-ray emission from different parts of the galactic plane (Aharonian et al. 2008a). The above results, as well as the prospects of continuous monitoring of a significant part of the sky, which might lead to exciting discoveries of yet unknown VHE transient phenomena in the Universe, justifies the new proposals of construction of large field-of-view high-altitude EAS detectors like *HAWK* (see for a review Aharonian et al. (2008a)). The sensitivity of *HAWK* in the energy region around 1 TeV is expected to be comparable to the sensitivity of *Fermi* around 1 GeV. In this regard *HAWK* will be complementary to *Fermi* for continuous monitoring of more than 1 sr fraction of the sky over three decades of energy from 100 GeV to 100 TeV.

The remarkable success of observational gamma-ray astronomy in the GeV and TeV energy regimes, together with recent intensive theoretical and phenomenological studies of acceleration and radiation processes, supply a strong rationale for the further exploration of the sky at high and very high energies. Although generally the main motivations of gamma-ray astronomy remain unchanged, the recent observational results have introduced important corrections to our understanding of many relevant phenomena. They revealed new features which in many cases require revisions of current theoretical models and formulations of new concepts. Hopefully the operation of *Fermi* over the next several years will be accompanied by observations with more sensitive next generation ground-based detectors, in particular CTA and *HAWK*. The data

obtained in the enormous energy range from 100 MeV to 100 TeV will provide deeper insight into a number of fundamental problems of high-energy astrophysics. Some of these topics are discussed in this chapter in the context of the role of VHE gamma-rays in the understanding of the origin of galactic cosmic rays.

3 Supernova Remnants and Galactic Cosmic Rays

Gamma-ray astronomy has a key role to play in solving the problem of origin of galactic cosmic rays. The basic concept formulated by pioneers of the field in the 1950s and 1960s (see, e.g., Morrison (1957), Ginzburg and Syrovatskii (1964)) is simple and concerns both the acceleration and propagation aspects of cosmic rays. Namely, while the localized gamma-ray sources exhibit the sites of production/acceleration of cosmic rays, the angular and spectral distributions of the diffuse galactic gamma-ray emission provide information about the character of propagation of relativistic particles in galactic magnetic fields.

The energy spectrum of cosmic rays has two distinct features – the so-called knee and ankle around 10^{15} and 10^{18} eV, respectively, with a cutoff above 10^{20} eV. While there is little doubt that all particles below the *knee* are of galactic origin, the extremely high energy cosmic rays above the *ankle* most likely are produced outside the Galactic Disk. The origin of the energy interval between the *knee* and *ankle* is less clear.

The acceleration, accumulation, and effective mixture of relativistic particles, through their convection and diffusion in galactic magnetic fields, results in a formation of the so-called sea of galactic cosmic rays. The level of this “sea” is contributed by operation of all galactic sources over a relatively long time, determined by the cosmic ray confinement time in the Galaxy of $T \approx 10^7$ year (e.g., Berezhinsky et al. (1990), Gaisser (1990)). Unless this level is not significantly different from the directly measured (local) cosmic ray flux, the cosmic ray energy density in the Galactic Disk should be close to 1 eV cm^{-3} . Because of the relatively steep energy spectrum of cosmic rays, $N(\geq E) \propto E^{-1.6}$, more than 90% of this density is contributed by particles with energy less than 100 GeV.

The cosmic ray production rate in the Galaxy can be reliably estimated using direct measurements of the local flux of cosmic rays as well as the secondary-to-primary ratio which contains information about the confinement time of cosmic rays in the Galaxy. Such an estimate is based on the assumption that cosmic rays are homogeneously distributed in the Galaxy, otherwise it is rather insensitive to the details characterizing the confinement region (density, volume, etc.). Namely, both the disk and halo models of galactic cosmic rays give similar, with an uncertainty of a factor of 3, production rate of cosmic rays in our Galaxy (see, e.g., Berezhinsky et al. (1990), Gaisser (1990))

$$\dot{W}_{\text{CR}} = (0.3 - 1) \times 10^{41} \text{ erg s}^{-1}. \quad (15.1)$$

Supernova Remnants (SNRs) are generally believed to be the principal contributors to the bulk of the observed cosmic ray flux. This idea has been proposed in early 1930s by Baade and Zwicky (1943). The main *phenomenological* argument in favor of this hypothesis is based on the fact that the luminosity of galactic cosmic rays given by (15.1) can be supported by SNRs if a quite reasonable fraction, 10% or so, of the total mechanical energy of SN explosions in our Galaxy is eventually released in cosmic rays. This is, of course, a strong but not yet a decisive argument, given that other nonthermal galactic source populations, in particular pulsars and

their wind nebulae, X-ray binaries, young stars with powerful mechanical winds, also meet, at least formally, this energy requirement.

The second key argument in favor of SNRs as sources of galactic CRs has a more theoretical background/motivation, namely, it relies on the diffusive shock acceleration (DSA) paradigm applied to young SNRs. Over the last 20 years the basic properties of this model have been comprehensively studied and cross-checked using different computational approaches. In particular, it has been realized that in effective accelerators the shocks are modified by the pressure of accelerated particles, and that this nonlinear effect should have a strong impact on the formation of the energy distribution of accelerated particles (for a review, see, e.g., Malkov and Drury (2001)). The nonlinear regime of diffusive shock acceleration implies (by definition) very high, as large as 50%, efficiency of energy release in cosmic rays. Moreover, the nonlinear effects predict strong, by an order of magnitude, amplification of magnetic field, e.g., due to nonresonant streaming instability of charged energetic particles propagating through the ambient plasma (Bell 2004). Note that the high efficiency of transformation of mechanical energy to nonthermal particles, $\kappa \geq 10\%$, and large magnetic field of order of $100 \mu\text{G}$ or more, are two *necessary conditions* for explanation of galactic cosmic rays by SNRs up to the “knee” around 10^{15} eV. In this regard, the nonlinear shock acceleration is a key element within the SNR paradigm of galactic cosmic rays.

The three principal factors – (1) the high efficiency of transformation of the available kinetic energy of bulk motion into nonthermal particles, (2) the hard energy spectra of protons extending to multi-TeV energy region, and (3) the relatively high density of the ambient gas – make the young supernova remnants viable sources of gamma-rays resulting from production and prompt decay of secondary π^0 -mesons. Thus, the most straightforward test of acceleration of proton and nuclei in SNRs can be performed via search for π^0 -decay gamma-rays – either directly from shells of young SNRs (Drury et al. 1994) or from nearby clouds interacting with an expanding SNR shell (Aharonian et al. 1994). It has been argued that the TeV gamma-ray domain is the best energy band to explore this possibility – from the point of view of the superior performance of the detection technique and because of the key information about particle acceleration carried by TeV gamma-rays (Drury et al. 1994). The nonlinear version of DSA allows definite observational predictions. The distinct feature of this model is the shape of the energy spectrum of accelerated particles which deviates from the standard power-law-type distribution. It has a rather concave curvature; at low (GeV) energies the spectrum is relatively steep with differential spectral index larger than 2, but at highest energies the spectrum becomes very hard. In the case of strongly modified shocks, the spectrum just before the high-energy cutoff can be as hard as $E^{-1.5}$ (Malkov and Drury 2001). These features are reflected in the nonthermal emission of accelerated electrons and protons.

These theoretical predictions initiated significant efforts toward detection of TeV gamma-rays from SNRs using the first-generation imaging Cherenkov telescopes of the *Whipple*, *HEGRA*, and *CANGAROO* collaborations. The first negative results reported in the late 1990s have been promptly interpreted by some experts as a failure of SNRs in general, and DSA in particular, to be responsible for the production of galactic CRs. However, given the limited sensitivity of the first-generation detectors, as well as large uncertainties in several key model parameters, in particular of the ratio n/d^2 (where n is the density of the ambient gas, and d is the distance to the source), these conclusions were premature, and somewhat exaggerated. As clearly stated by Drury et al. (1994), only stereoscopic arrays of atmospheric Cherenkov telescope with adequate sensitivity and morphological and spectrometric capabilities can provide meaningful probes of TeV gamma-ray production in SNRs. Later, gamma-ray emission has

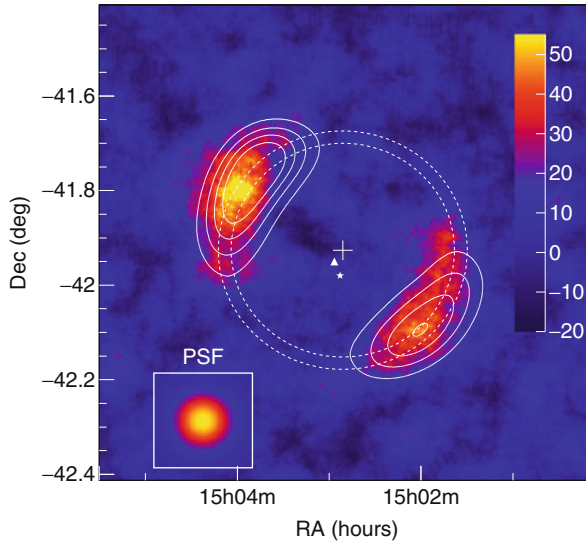
indeed been detected from several SNRs (see for a review Aharonian et al. (2008a)). Presently six young shell type SNRs – Cas A, RX J1713.7-3946, RX J0852-4622 (Vela Jr), RCW 86, SN 1006, Tycho – are identified as TeV gamma-ray emitters.¹ These objects are not strong GeV gamma-ray emitters. So far low gamma-ray fluxes are reported by *Fermi* LAT only from two young SNRS – Cas A (Abdo et al. 2010b) and RX J1713.7-3946 (Abdo et al. 2011). Mid-age SNRs appear to be more prolific gamma-ray emitters at low energies. Several such objects have been reported by the *Fermi* LAT and *AGILE* groups. Some of them, e.g. W28 and IC 433 have also TeV counterparts as reported by the HESS, MAGIC and VERITAS collaborations. They are, most likely, SNR/molecular-cloud interacting systems. These extended gamma-ray structures are most likely linked to the SNR/molecular-cloud interacting systems. Finally, several galactic TeV gamma-ray sources spatially coincide with the so-called composite supernova remnants, objects with characteristic features of both standard shell-type SNRs and PWNe (plerions). At least in one case, the association of a TeV source with the composite SNR G0.9+0.1 is clearly established.

The number of SNRs showing shells in TeV gamma-rays slowly increases. However, it is not sufficient for effective population studies. Moreover, it is not obvious that with an increase of number of TeV gamma-ray-emitting SNRs, the population studies would lead to definite and unbiased conclusions. Both the particle acceleration and radiation processes are quite sensitive to the parameters which characterize the initial conditions of the SN explosion, as well as to the parameters of the ambient medium. Although a common feature of these objects is the young age and, consequently, the high shock speeds exceeding $2,000 \text{ km s}^{-1}$ (one of the key conditions for acceleration of particles to TeV energies), the multiwavelength properties of these objects are quite different. This can be explained by many factors, in particular by the initial power of SN explosions as well by the location of these objects in environments with essentially different parameters. Therefore the most effective approach to these objects seems to be studies of the origin of relativistic particles and their radiation mechanisms on the source-by-source basis. Three distinct examples of young SNRs are discussed below.

4 SN 1006: A Classical Shell-Type Supernova Remnant

This object is the remnant of an extremely bright event recorded in 1006AD (Stephenson and Green 2001). SN 1006 was the first SNR from which a nonthermal component of X-rays has been revealed and identified with synchrotron radiation of multi-TeV electrons (Koyama et al. 1995). Motivated by this finding which implies acceleration of electrons in the shell to multi-TeV energies, the *CANGAROO* collaboration has conducted extensive observations and soon reported a detection of TeV gamma-ray signal from this young SNR. However, subsequent observations of the source in 2003–2004 with the *HESS* telescope array did not confirm this result; the flux upper limit set by *HESS* was an order of magnitude below the flux claimed by the *CANGAROO* collaboration. Interestingly, the *HESS* upper limit was in a good agreement with

¹Remarkably, in the survey of the galactic plane conducted with the *HESS*, an extended source, HESS 1731-374 with a shell-type structure has been found with no clear counterpart at other wavelengths. However, soon a new shell-type SNR was reported, both in radio and nonthermal X-rays, positionally coinciding with the gamma-ray source. If this association is correct, this would make HESS 1731-374 as the first SNR discovered in gamma-rays (Abramowsky et al. 2011).

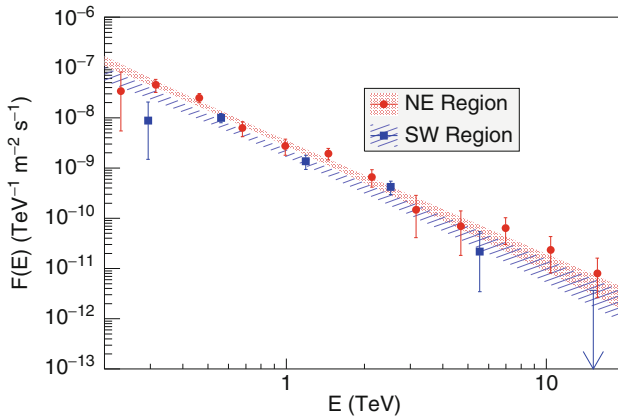


■ Fig. 15-1

HESS γ -ray image of SN 1006 (Accero et al. 2010). The *white contours* correspond to a constant X-ray intensity as derived from the XMM-Newton flux map and smoothed to the HESS point spread function. The *white cross*, *star*, and *triangle* show the center of the X-ray image obtained in different ways. The *inset* shows the HESS PSF using an integration radius of 0.05°

conservative theoretical expectations for both hadronic and leptonic components of gamma-rays, given the low-density environment around SN 1006 (the source is 500 pc above the galactic plane), and the large magnetic field which follows from the small-scale filamentary structure of the nonthermal X-ray emission (Bamba et al. 2003; Völk et al. 2003). Later, more than 100 h exposure of this source by the HESS telescope array resulted in the detection of a statistically significant gamma-ray signal (Accero et al. 2010). The gamma-ray emission is concentrated in two extended regions in North-East (NE) and South-West (SW) showing bipolar morphology which correlates with the image of nonthermal X-rays (see ► Fig. 15-1). Since in this object we deal with a quite homogeneous environment, such a morphology indicates the sites of concentration of highest energy particles. It has been argued that a possible reason for the bipolar picture could be the dependence of efficiency of injection of suprathermal particles on the angle between the ambient magnetic field and the shock normal (Völk et al. 2003), but this is a rather complex problem which needs deep theoretical studies.

The energy spectra of the NE and SW regions are shown in ► Fig. 15-2. The integral flux above 1 TeV corresponds to less than 1% of the Crab flux. The energy spectra from both regions are compatible with power-law distributions, $F(E) \propto E^{-\Gamma}$, with a photon index $\Gamma \approx 2.3$. If one relates the slope to the power-law index of accelerated protons, then it is somewhat steeper than the proton spectra generally expected from the diffusive shock acceleration theory. However, the slope can be more naturally interpreted as a result of the “nominal” E^{-2} type gamma-ray spectrum combined with an *early* high-energy-cutoff around 10 TeV. In particular the gamma-ray spectrum can be explained by interactions of accelerated protons with the ambient gas, assuming energy distribution of protons in the form $E_p^{-\alpha} \exp(-E/E_0)$ with $E_0 \sim 80$ TeV (Accero et al. 2010).



■ Fig. 15-2

Differential energy spectra of SN 1006 extracted from the two regions NE and SW. The shaded bands correspond to the range of the power-law fit, taking into account statistical errors

Moreover, in the framework of *hadronic* models, the spectrum of protons flatter than E_p^{-2} is a quite robust condition opposed by the available energy budget of the source. Indeed, given that the VHE emission is concentrated in two relatively compact regions of the shell, for the gas density of the ambient medium $n \sim 0.1 \text{ cm}^{-3}$ and the distance to the source $\approx 2 \text{ kpc}$, the total efficiency of acceleration of protons to explain the observed gamma-ray flux already exceeds 50% of the total mechanical energy of the SN explosion, $E_{\text{SN}} = 1.4 \times 10^{51} \text{ erg}$ (Accero et al. 2010). Note, however, that for the nebular magnetic field $B \leq 100 \mu\text{G}$, a major fraction of reported TeV gamma-rays can be contributed by the Inverse Compton (IC) scattering of electrons. Since the 2.7 K Cosmic Background Radiation (CMBR) serves as the main target for IC scattering, the flux of IC gamma-rays depends, for the given synchrotron X-ray flux, only on the strength of the ambient magnetic field. In particular, within the simple one-zone model the interpretation of the observed TeV gamma-rays entirely by the IC component requires $B \approx 30 \mu\text{G}$ (Accero et al. 2010). Note that, independent of the origin of gamma-rays, this value should be considered as a robust lower limit to the average magnetic field; smaller fields would lead, for the given X-ray flux, to overproduction of IC gamma-rays ($F_\gamma \propto 1/B^2$). On the other hand, the magnetic field can be somewhat larger, e.g., by 50%. In this case the deficit of IC gamma-rays (by a factor of 2) can be compensated by the contribution from p-p interactions. Finally, larger magnetic fields, $B \geq 100 \mu\text{G}$, would imply almost pure hadronic origin of the detected TeV gamma-ray emission. Although the one-zone model is a simplification of the gamma-ray production scenario (Aharonian and Atoyan 1999), the above estimates of the average magnetic field in the shell describe quite correctly the contributions of the leptonic and hadronic components to the TeV gamma-radiation of SN 1006.

The identification of the origin of gamma-rays is a key issue which should allow us to fix with a good accuracy the average magnetic field in the shell and derive the total energy released in accelerated electrons and protons. This general statement is relevant to all gamma-ray emitting SNRs, but SN 1006 is a special case, given an almost perfect combination of conditions in this object for realization of effective particle acceleration. This concerns, in particular, the “optimal age” of the source, the large energy of SN explosion, $E_{\text{SN}} \approx 1.4 \times 10^{51} \text{ erg}$, and the speed

of the shock wave of order of $\approx 5,000 \text{ km s}^{-1}$ of the expanding in a low-density homogeneous environment not crowded with sources. On the other hand the gamma-ray flux of SN 1006 is very low, and its detailed morphological and spectroscopic studies will be possible only with the next generation of ground-based instruments like CTA. The extension of the spectral coverage to both $\leq 100 \text{ GeV}$ and $\geq 10 \text{ TeV}$ energies should be one of the principal issues in the context of separation of leptonic and hadronic components of gamma-radiation.

Meanwhile, the currently available VHE gamma-ray data on SN 1006 already contain quite important information about the source. As mentioned above, the ratio of X-ray and gamma-ray fluxes tells us that the average magnetic field in the shell cannot be less than $30 \mu\text{G}$. This implies that the energy of electrons cannot significantly exceed $3 \times 10^{47} \text{ erg}$ (Accero et al. 2010). On the other hand any significant contribution of the hadronic component to gamma-rays would be possible if the total energy in protons exceeds 3×10^{50} . This implies that the proton-to-electron ratio is larger than 10^3 , i.e., an order of magnitude exceeds the ratio observed in local cosmic rays. This ratio could be quite different (larger or smaller) in the case of other SNRs. Two distinct representatives of young SNRs with essentially different features are discussed below.

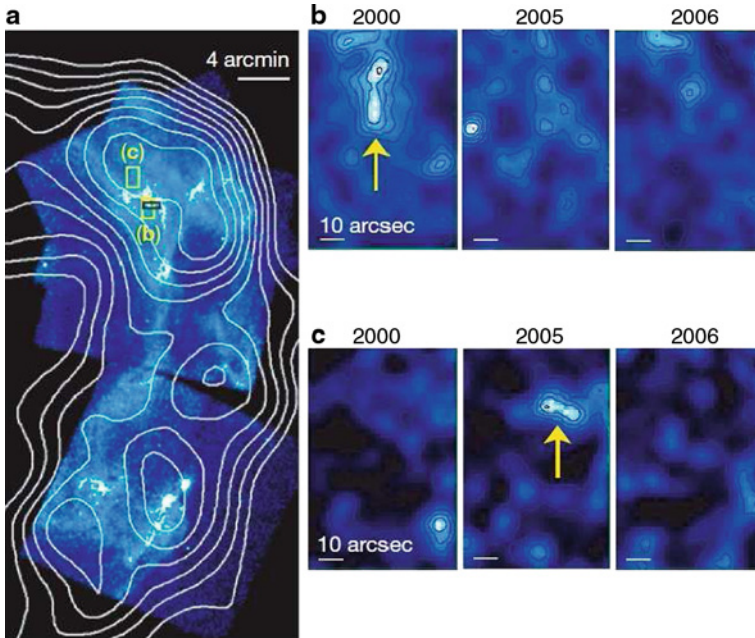
5 RX J1713.7-3946: A Unique SNR

RX J1713.7-3946, an extended, approximately 1° diameter X-ray source, is a unique (atypical) representative of young supernova remnants. While the synchrotron radio emission and thermal X-rays are two distinct components of shell-type SNRs, RX J1713.7-3946 shows weak radio emission, and no thermal X-ray emission at all. On the other hand, this object is a powerful nonthermal X-ray and TeV gamma-ray emitter.

The spectral, temporal, and morphological properties of nonthermal X-ray emission of this source provide useful information about conditions in sites of particle accelerators. While the only viable interpretation of the nonthermal X-ray emission is the synchrotron radiation, the diffusive shock acceleration seems to be the most likely mechanism that boosts the electrons to multi-TeV energies. The synchrotron spectrum extends to 10 keV and beyond, which is an indicator that the electron acceleration in this source proceeds in the extreme Bohm diffusion regime (Tanaka et al. 2008). The deep observations of the source with sub-arcsecond resolution of the *Chandra* X-ray telescope resulted in the detection of a complex network of bright X-ray filaments and knots (Lazentic et al. 2004; Uchiyama et al. 2003), as well as in the discovery of X-ray variability of compact ($\leq 10^{17} \text{ cm}$) regions (Uchiyama et al. 2007) on few-year timescales (see [Fig. 15-3](#)). A likely explanation of these features is the significant, by a factor of 10 or more, enhancement of the magnetic field compared to the interstellar magnetic field. The large shock speed, $v \sim 3,000 \text{ km s}^{-1}$, the high turbulence allowing particles to diffuse in the Bohm regime, and the large magnetic field, $B \geq 100 \mu\text{G}$, are three necessary conditions for acceleration of particles in SNRs beyond 100 TeV (Lagage and Cesarsky 1983). This makes RX J1713.7-3946 an attractive target for high-energy gamma-ray observations.

Initially, RX J1713.7-3946 was reported as a source of TeV gamma-rays by the CANGAROO group. The *HESS* observations confirmed this result, and, more importantly, revealed a nice TeV gamma-ray image of the remnant (Aharonian et al. 2004) shown in [Fig. 15-4](#). The overall shell structure and its correlation with the nonthermal X-ray image is clearly visible.

The energy spectrum of gamma-rays integrated over the entire remnant (Aharonian et al. 2007) is shown in [Fig. 15-5](#). It extends over two decades in energy, from 300 GeV to $\geq 30 \text{ TeV}$.



■ Fig. 15-3

The filamentary structure and variability of X-ray emission observed by the Chandra X-ray satellite from the western shell of RX J1713.7-3946 (Uchiyama et al. 2007). (a): *Left panel*: A Chandra X-ray mosaic image overlaid with TeV gamma-ray contours from HESS observations. For the distance to the source $d = 1$ kpc, the angular size 1 arcmin corresponds to 0.29 pc. The Chandra images are obtained for the energy interval 1–2.5 keV, with a pixel size of 2 arcsec, and smoothed with Gaussian kernel of 8 arcsec. *Right panel*: (b): A sequence of X-ray observations in July 2000, July 2005, and May 2006 for the small region labeled as (b) in the *Left panel*. A well-defined emission feature in the image from 2000 faded before the year 2005, revealing variability on timescales of years. (c): *Hard-band* (3.5–6 keV) image of the box labeled as (c) in *Left panel*. The hot spot indicated by a yellow arrow appeared in the 2005 July image and faded before May 2006

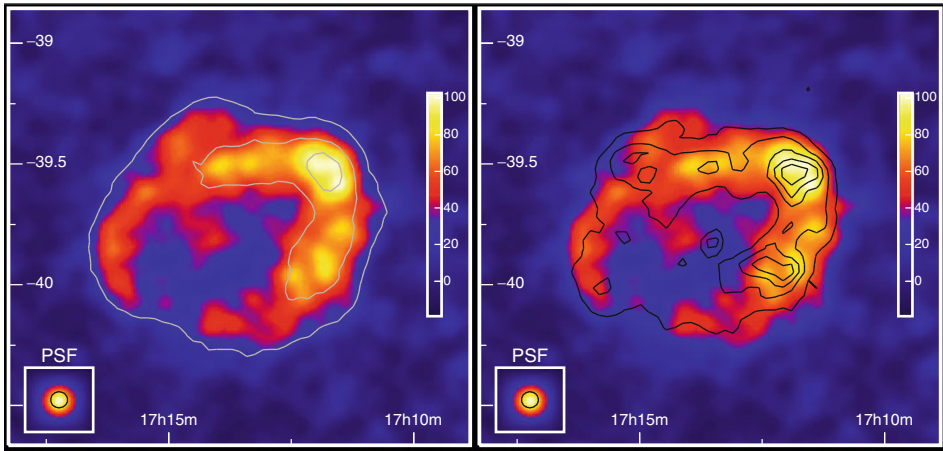
Note that the data above 30 TeV show a statistically significant (4.8σ) signal. The highest energy spectral points above 10 TeV are crucial for the conclusion whether the energy spectrum contains a break or it has a sharper behavior, e.g., an exponential cutoff. The present data are well described (Aharonian et al. 2007) either by a “broken power-law,”

$$dN/dE \propto (E/E_B)^{-\Gamma_1} [1 + (E/E_B)^{1/a}]^{a(\Gamma_1 - \Gamma_2)}, \quad (15.2)$$

with $a = 0.6$, $\Gamma_1 = 2.00 \pm 0.05$, $\Gamma_2 = 3.1 \pm 0.2$, and $E_B = 6.6 \pm 2.2$ TeV, or by a power-law spectrum with an exponential cutoff, written in a general form

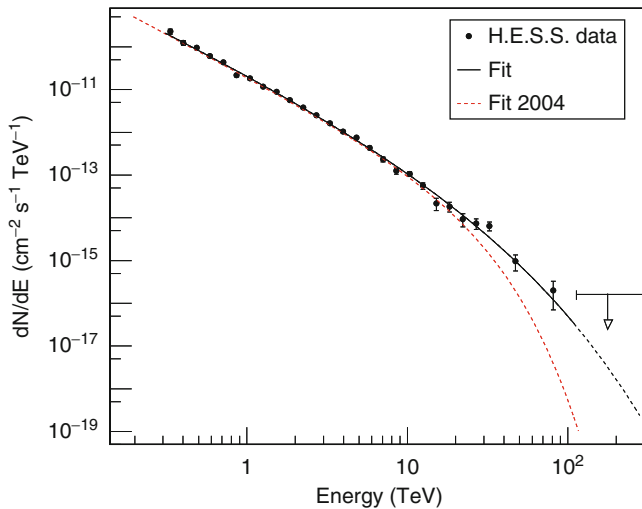
$$dN/dE \propto E^{-\Gamma} \exp\left[-(E/E_0)^\beta\right], \quad (15.3)$$

In this presentation the parameters Γ , β , and E_0 correlate with each other, thus good fits can be obtained for different combinations, in particular for (1) $\Gamma = 1.79 \pm 0.06$, $E_0 = 3.7 \pm 1.0$, $\beta = 0.5$, and (2) $\Gamma = 2.04 \pm 0.04$, $E_0 = 17.9 \pm 3.3$, $\beta = 1.0$.



■ Fig. 15-4

The gamma-ray image of RX J1713.7-3946 obtained with the HESS telescope array (Aharonian et al. 2007). Shown is an acceptance-corrected excess count (statistical significance more than 50 standard deviations) smoothed with a Gaussian kernel of 2 arcmin. On the *left-hand side*, the overlaid *light-gray* contours illustrate the significance of the different features. The levels are at 8, 18, and 24σ . On the *right-hand side*, the *black lines* show the 1–3 keV nonthermal X-ray contours of the source

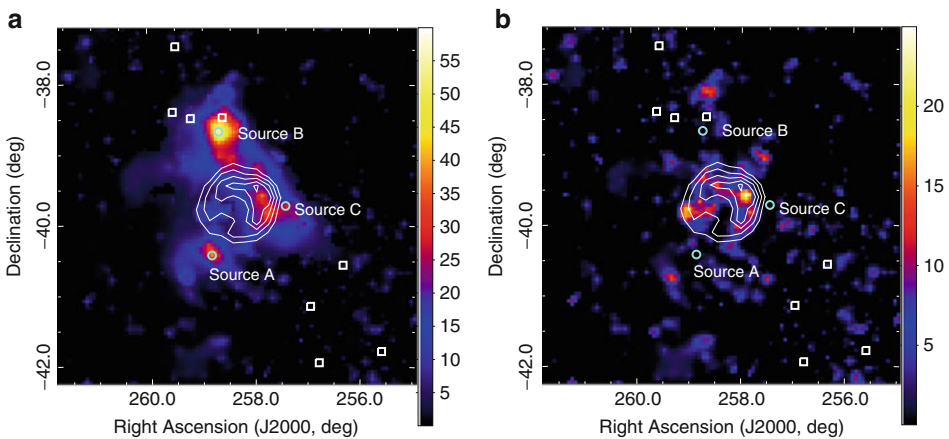


■ Fig. 15-5

The energy spectrum of gamma-rays from RX J1713.7-3946 (Aharonian et al. 2007). The error bars imply $\pm 1\sigma$ statistical uncertainty. The *black solid line* corresponds to fit given by (15.3) with $\beta = 0.5$, $\Gamma = 1.8$, and $E_c = 3.7$ TeV. The extrapolation of this spectrum beyond the fit range (*dashed*) is shown for illustration. A model-independent upper limit in the energy range from 113 to 300 TeV is also shown. The *dashed red line* corresponds to the fit reported initially by the HESS collaboration based on the statistically limited 2004 data set (Aharonian et al. 2004)

The presentation given by (☛ 15.3) is of a special interest. In particular, in the framework of hadronic models it reflects the energy spectrum of shock-accelerated protons which can be presented in a similar form, $\propto E_p^{-\alpha} \exp[-(E_p/E_{p,0})^s]$. The parameters characterizing the spectra of protons and secondary gamma-rays are related as $E_0 \sim 1/20E_{p,0}$, $\beta \approx 2s$, and $\Gamma \approx \alpha - \delta\alpha$ with $\delta\alpha \approx 0.05 - 0.1$ (Kelner et al. 2006). When the particle acceleration proceeds in the most effective (Bohm-diffusion) regime, the spectrum of protons not suffering radiative losses is described by a simple exponential cutoff, i.e., $s = 1$. Thus the fit of the measured gamma-ray spectrum by (☛ 15.3) with $\beta = 0.5$ would imply that protons are accelerated in the regime close to the Bohm diffusion with a power-law index $\alpha \approx \Gamma + 0.05 \approx 1.85$ and cutoff energy $E_{p,0} \approx 20E_0 \approx 75$ TeV. However one should note that the available gamma-ray data do not allow derivation of a unique proton spectrum. Instead we can only state that within the energy range 0.3–300 TeV the proton spectrum in SNR RX J1713.7-3946 is well described by a power-law with an index α within 1.7–2 and a cutoff/transition region $E_{p,0}$ between 30 and 100 TeV (Vilante and Vissani 2007). Moreover, the possible non-negligible contribution of IC gamma-rays to the overall gamma-ray flux makes the derivation of the proton spectrum more complex and requires detailed *spatially resolved* energy spectra of TeV gamma-rays in a broad dynamical range, from MeV/GeV to multi-TeV energies.

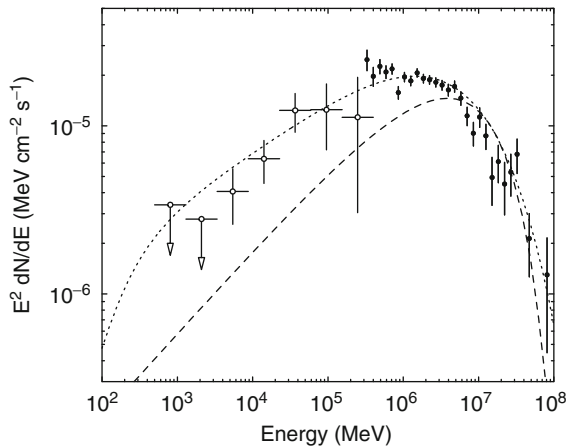
In this regard, the recently reported spectrum of the source by the *Fermi* collaboration (Abdo et al. 2011) is an important step toward understanding the origin of nonthermal emission of this source. Unfortunately, the quality of the *Fermi* data especially at energies around and below 1 GeV at which the angular resolution of the instrument exceeds the size of the source is not compatible with the X-ray and TeV gamma-ray data. Although a clear signal above 500 MeV has been detected by *Fermi* with a location and size that matches the TeV source, the shape of the image significantly depends on the background model (see ☛ Fig. 15-6). This introduces a significant uncertainty when we want to compare the fluxes and energy spectra at GeV and TeV energies.



☛ Fig. 15-6

The images of gamma-rays of RXJ 1713.7-3946 obtained with *Fermi* assuming two different models for the background. Only events above 500 MeV have been used. *HESS* TeV emission contours are shown in white (Abdo et al. 2011)


The immediate conclusion to be drawn from the GeV-TeV spectral energy distribution (SED) shown in [Fig. 15-7](#) is that the emission in the GeV energy band is somewhat suppressed. This does not agree with the extrapolation of the TeV gamma-ray spectrum given in the form of ([15.2](#)) or ([15.3](#)) with a power-law index close to $\Gamma = 2$. This implies that within the framework of one-zone hadronic models the proton spectrum should be harder than E^{-2} . In the *Fermi* collaboration paper this straightforward conclusion has been overinterpreted and generalized with a claim that the *Fermi* data exclude the hadronic origin of gamma-radiation (Abdo et al. 2011). However, one should note that although the hadronic models of gamma-radiation of this source do indeed face certain problems (see below), the *Fermi* data cannot be treated as a strong argument against the hadronic origin of gamma-radiation; there are several caveats which make this claim quite shaky and premature. First of all, the SED shown in [Fig. 15-7](#) assumes that gamma-rays detected at the GeV and TeV energies are coming from the exactly same regions. However, this assumption is not supported by the current data. One cannot exclude, e.g., that the flux of the TeV gamma-ray emission from the GeV “hot spots” in the *Fermi* map is less, e.g., by a factor of 2, than the overall TeV flux (see [Fig. 15-6](#)). If so, it would be more appropriate to compare the TeV gamma-ray flux extracted from the regions of GeV excess. Secondly, even assuming that the reported GeV and TeV gamma-ray fluxes arrive from the same regions, one can easily match the GeV and TeV spectral points by adopting a hard spectrum of protons with a power-law index $\Gamma \leq 1.8$ (see [Fig. 15-7](#)). Generally, such a hard proton spectra cannot be



■ Fig. 15-7

The spectral energy distribution of GeV and TeV gamma-rays from RX J1713.7-3946 based on the *Fermi* (Abdo et al. 2011) and *HESS* (Aharonian et al. 2007) data. The “hadronic” (due to p-p interactions) and “leptonic” (due to IC) gamma-ray spectra calculated within a simple one-zone model are also shown. The corresponding curves 1 and 2 are from Figs. 14 and 21 of Tanaka et al. (2008). The IC curve is obtained for the electron spectrum derived from the synchrotron X-ray flux assuming magnetic field of about $16 \mu\text{G}$. The π^0 -decay gamma-ray spectrum corresponds to the power-law spectrum of protons with $\Gamma = 1.7$. Note that the GeV and TeV gamma-ray production regions can only partly overlap, thus the interpretation of GeV and TeV measurements within one-zone models could lead to misleading conclusions

excluded; moreover they are anticipated by nonlinear acceleration models. Some other natural ways to suppress the GeV fluxes in hadronic models of this source are discussed below.

From  Fig. 15-7 one can see that the GeV points are above the flux predicted by a one-zone inverse Compton model. Again, this cannot be treated as an argument against the leptonic models, but rather would imply, e.g., the presence of the second (low-energy) component of electrons in the supernova remnant (Tanaka et al. 2008). In summary, the recent *Fermi* results do provide a very important and complementary information to the data obtained in the TeV band, but do not robustly constrain either hadronic or leptonic models suggested for this source. This issue is discussed in more details in the next section.

The likely distance to the SNR RX J1713.7-3946 of about 1 kpc favors a young age of the source between 1,000 and 3,000 years. If so, RX J1713.7-3946 can be the result of the supernova explosion which was registered in the Chinese historical records (Wang et al. 1997) in 393AD as a guest star. Therefore RX J1713.7-3946 can be formally treated as a representative of young galactic SNRs like SN 1006, Vela Junior, Tycho, Kepler, Cas A. However, RX J1713.7-3946 seems to be a unique object with quite unusual characteristics. This circumstance makes the identification of the gamma-ray mechanism, and hence the nature of the parent particles (electrons or protons?) a nontrivial problem. Meanwhile the question of the origin of gamma-rays in the context of the popular for these days debates “hadronic versus leptonic models” is not a matter of purely academic interest, but has a more fundamental implication related to the so-called SNR paradigm of galactic cosmic rays which postulates the major role of SNRs to the production of galactic cosmic rays.

5.1 Challenges of Hadronic Models

One of the most puzzling features of RX J1713.7-3946 is the lack of thermal X-ray emission which has been claimed to be a decisive argument against the hadronic models of this source (see, e.g., Katz and Waxman (2008)). The tight upper limit on the thermal X-ray flux of RX J1713.7-3946 is explained by the supernova explosion inside a wind-blown bubble with a very low gas density, $n_{\text{gas}} \ll 1 \text{ cm}^{-3}$ (Cassam-Chenai et al. 2004; Slane et al. 1999; Tanaka et al. 2008). Formally, the very low-density ambient gas favors the leptonic origin of TeV gamma-rays produced through the inverse Compton scattering of ultrarelativistic electrons which are responsible also for the nonthermal synchrotron X-ray emission. However, the hadronic origin of gamma-rays in a low-density environment cannot be a priori rejected. Moreover, this scenario could be, to a certain extent, a quite attractive option, in particular in the context of the “SNRs-GCRs” paradigm. Indeed, within the hadronic models of gamma-ray production, the low gas density is accommodated in expense of corresponding increase of the nonthermal energy budget in protons ($W_p \propto F_\gamma n_{\text{gas}}^{-1}$). This would imply an extremely effective acceleration of cosmic rays when the major fraction of the kinetic energy of explosion is converted into accelerated particles. Remarkably, the recent developments of the nonlinear theory of DSA not only allow, but, in fact, demand particle acceleration with efficiency significantly exceeding the “standard” value of 10% (see below). Such an effective acceleration most likely does not take place in all SNRs, and in this regard the major contribution to the galactic cosmic rays perhaps is provided by a subclass of SNRs with reduced power of thermal emission of the shocked gas. In any case, it is likely that VHE gamma-ray production efficiency of SNRs anticorrelates with the thermal X-ray emission luminosity. In particular, the reported weak gamma-ray emission

(or its absence) from three archetypical representatives of young SNRs with powerful thermal X-ray emission, Tycho, Kepler, and Cas A, leads the upper limit on the total energy content of accelerated TeV protons and nuclei: $W_p \leq 10^{49}$ erg. The high X-ray emission implies existence high-velocity ($\geq 3,000$ km s $^{-1}$) shocks and high-density plasma, both being crucial components for effective production of TeV gamma-rays of hadronic origin. Quite paradoxically, we see very low gamma-ray fluxes from the above-mentioned “right age” SNRs (in the sense of ability of acceleration of multi-TeV particles) and, at the same time, a very strong TeV gamma-ray emission shows the peculiar SNR RX J1713.7-3946.

In SNR shocks the plasma is heated up to a temperature $kT = \frac{3}{16} m_p v_{sh}^2$, where m_p is the proton mass, and v_{sh} is the shock speed. For the typical shock speed in young SNRs exceeding 2,000 km s $^{-1}$, the temperature can be as high as 10 keV. Generally, the high proton temperature does not necessarily imply high electron temperature. The exchange of energies between the electron and proton components is realized via Coulomb collisions, which, however, for the conditions typical for SNRs are not sufficiently effective to establish electron–proton equipartition. On the other hand, X-ray observations show that electrons in young SNRs are heated to keV temperatures. This is explained by a hypothetical mechanism which in collisionless shocks operates, e.g., through excited plasma waves.

The low emissivity of thermal X-ray emission of RX J1713.7-3946 implies low plasma density $n \leq 0.1$ cm $^{-3}$ and/or low electron temperature ($kT_e \leq 0.1$ keV). The latter can be explained, in principle, by low proton–electron exchange rate. However, the conservative estimates show that in a standard regime of particle acceleration in RX J1713.7-3946, the electrons of ambient plasma are heated to quite high temperatures, even when the heating proceeds only through the Coulomb exchange (Zirakashvili and Aharonian 2010; ?). Thus, the problem of low electron temperature cannot be reduced merely to the problem of electron–proton energy exchange rate. A more relevant reason could be that a large fraction of energy goes into the cosmic-ray production due to the nonlinear effects in the shock, and thus protons appear to be under-heated. A certain support of this conclusion could be the recent measurements of the shock speed and the postshock proton temperature in a young ($\approx 2,000$ year) supernova remnant RCW 86 (Helder et al. 2009). The optical spectroscopic measurements of thermal Doppler broadening of hydrogen lines revealed a postshock proton temperature in this object $kT_p = 2.3 \pm 0.3$ keV (Helder et al. 2009). On the other hand, the estimate of the shock speed by the same authors based on the measurement of the proper motion of the shock (from the comparison of two X-ray images taken by *Chandra* in 2004 and 2007) appeared extremely large, $(6.0 \pm 2.0) \times 10^3$ km s $^{-1}$. If this estimate is correct, its most exciting implication is that according to standard shock relations one should expect proton temperature by an order of magnitude higher than the measured one! While the high shock speed can be explained by the expansion of the remnant in a low-density cavity blown by the stellar wind, the very low proton temperature could be related to the very high efficiency of proton acceleration which converts the major, $f \geq 0.5$, fraction of the kinetic energy of explosion into nonthermal particles. Correspondingly the fraction of available energy which goes to the heating of the ambient plasma will be significantly reduced. The modifications of the SNR structure and dynamics by effective particle acceleration indeed allow suppression of the gas heating; the postshock gas temperature in principle can be as small as six times the temperature of the ambient gas (Drury et al. 2009), i.e., it can be suppressed, in principle, to the point where thermal X-ray emission is no longer expected. However, such a (hypothetical) situation can be realized only for very small Mach numbers (Vink et al. 2010) which most likely is not the case of RX J1713.7-3946. The detailed numerical calculations show

that in the case of RX J1713.7-3946 the thermal X-ray bremsstrahlung can be suppressed to the level of the upper limit on the thermal X-ray flux only if one assumes very low density of the ambient gas, $n = 0.02 \text{ cm}^{-3}$ (Zirakashvili and Aharonian 2010). However, the X-ray flux still is expected significantly higher due to the contribution from the X-ray line emission (Elisson et al. 2010; Zirakashvili and Aharonian 2010). Since within the hadronic models of gamma-rays, the limited budget of available energy of the source does not allow further reduction of plasma density, the overproduction of X-ray line emission can be avoided only assuming that the chemical abundance of heavy ions in X-ray-emitting regions is significantly reduced compared to the solar composition.

Within the hadronic models, gamma-rays are produced at interactions of accelerated protons, while nonthermal radio and X-ray components are due to synchrotron radiation of directly accelerated electrons. The hadronic models demand quite a large magnetic field of order of $100 \mu\text{G}$ or larger. This, for the given flux of synchrotron radiation, requires only $\approx 10^{46}$ erg in electrons. On the other hand, the total energy in protons should be 10^{50} erg or significantly larger if the background gas density $n \ll 1 \text{ cm}^{-3}$. Thus, in the hadronic scenario the electron-to-proton ratio is close to $K_{ep} = 10^{-4}$. This is two orders of magnitude smaller than the “standard” electron/proton ratio observed in local cosmic rays. This has been indicated (see, e.g., Katz and Waxman (2008)) as the second “trouble” for hadronic models. The ratio K_{ep} for a specific SNR of fixed age does not necessarily need, strictly speaking, to agree with the local cosmic ray value. Furthermore, the e/p ratio $\sim 10^{-2}$ measured in cosmic rays is relevant to the low-energy band, while the ratio $K_{ep} \sim 10^{-4}$ in RX J1713.7-3946 is derived for multi-TeV particles. Also, the low-energy electrons are produced at the later stages of SNR evolution, when the e/p ratio could be in principle different from the ratio at the early epochs of evolution of the remnant. Moreover, it is not obvious that the low-energy electrons and multi-TeV particles are contributed by the same source population. The issues related to radio-emitting electrons are more relevant to low-energy (GeV) protons in SNRs. In this regard, the measurements of low-energy gamma-rays from RX J1713.7-3946 with the *Fermi* LAT telescope will give us an important, model-independent information about the electron-to-proton ratio not distorted by radiative losses of electrons as well as by the escape of low-energy electrons and protons.

To conclude, the lack of thermal X-ray emission from RX J1713.7-3946 and the uncomfortably small ratio $K_{ep} \approx 10^{-4}$ do challenge the hadronic models of TeV gamma-rays, but yet cannot be considered as decisive arguments against hadronic models of gamma-rays of RX J1713.7-3946. They should be carefully addressed and explained in any detailed theoretical treatment of acceleration of protons and electrons in this unusual object.

Finally, the recent detection of weak and hard spectrum of the gamma-ray signal reported at multi-GeV energies has been interpreted as a “long-awaited” argument against the hadronic origin of gamma-rays (Abdo et al. 2011). This is, however, an apparent exaggeration of implications of the *Fermi* data. Indeed, the GeV gamma-ray “deficit” can be related to the hard ($E^{-1.8}$ or harder) spectrum of protons which generally is not prohibited by the DSA paradigm; moreover it is preferred by the nonlinear versions of shock acceleration models. Also, one should take into account that although the position and size of the GeV source match the TeV image, different parts of the shell can differently contribute to the observed GeV and TeV gamma-ray fluxes. If so, the interpretation of the total GeV and TeV gamma-ray fluxes integrated over the entire extended source within the simple one-zone models can be misleading. Note that in this source we have at least two emission zones related to the upstream and downstream of the forward shock. Moreover, if the reverse shock in this source plays a non-negligible role in the particle

acceleration, the number of zones of radiation will be doubled (Zirakashvili and Aharonian 2010). Finally, the hadronic gamma-rays can be produced in the dense clouds overtaken by the expanding shell. In this case we may anticipate comparable contributions from hadronic interactions, which take place in these dense formations, and from the inverse Compton scattering in the entire low-density shell. Such a “hybrid” model (Zirakashvili and Aharonian 2010) can readily match the *Fermi* and *HESS* points. While the overall shell can contribute to the region of intermediate (0.1–10 TeV) energies, gamma-rays from dense condensations can dominate in the low (GeV) and ultrahigh (above 10 TeV) energy intervals. It is interesting to note that because of slow diffusion, the relatively low-energy protons ($E \leq 100$ GeV) could not, during the age of the object, penetrate deep into the dense cores of these condensations. Since generally the diffusion has energy-dependent character, this effect should have less impact on highest energy protons. Consequently, we may expect suppression of GeV gamma-ray flux compared to the TeV flux. This effect discussed by Zirakashvili and Aharonian (2010) might provide an alternative explanation of the observed low ratio of GeV-to-TeV gamma-ray fluxes in this object.

5.2 Challenges of Leptonic Models

The major challenge for the leptonic models is the demand for a small magnetic field in the shocked shell of the remnant. The interpretation of the synchrotron and IC radiation components of relativistic electrons in RX J1713.7-3946 requires magnetic field around $15 \mu\text{G}$. Generally, the one-zone leptonic models fail to describe the spectral shape of gamma-rays below 1 TeV (see [Fig. 15-7](#)). The main problem of explanation of low-energy gamma-rays by IC component is related to the cooling break in the electron spectrum, and correspondingly to the position of the Compton peak which in the spectral energy distribution of gamma-rays appears above 1 TeV. Thus, the reduction of the break energy down to 200 GeV could in principle solve the problem. However, since the magnetic field in this model cannot significantly exceed $15 \mu\text{G}$, the only possibility to shift the Compton peak to sub-TeV energies is to assume that the remnant is much older than 10^3 years, which however is not supported by other properties of the source.

A more realistic solution for explanation of the overall gamma-ray spectrum, from 100 GeV to ≥ 10 TeV, seems the existence of an additional, low-energy component of electrons (Tanaka et al. 2008) which implies a deviation from the one-zone models. The latter should be considered as a convenient tool allowing simple qualitative estimates of some model parameters. However, it cannot be invoked for quantitative description of spectral properties of broadband emission of SNRs. Moreover, the DSA model implies, by definition, at least *two* distinct zones of nonthermal emission related to the downstream and upstream regions of the shock. In the case of formation of a reverse shock the number of zones is increased to four. This could be the case of RX J1713.7-3946, the radio and X-ray observations of which indicate to the possible existence of a second, *inner* shell with an angular radius of 0.25° , the half of the radius of the entire remnant. Within the multi-zone model, the constraints on the strength of the magnetic field are less robust, since the major contributions to the IC and synchrotron components of radiation can be formed in different regions, e.g., in the forward and inverse shocks with essentially different magnetic fields. The superposition of contributions from different zones results in quite complex morphological and spectral features of radiation.

6 DSA with Forward and Reversed Shocks Applied to RX J1713.7-3946

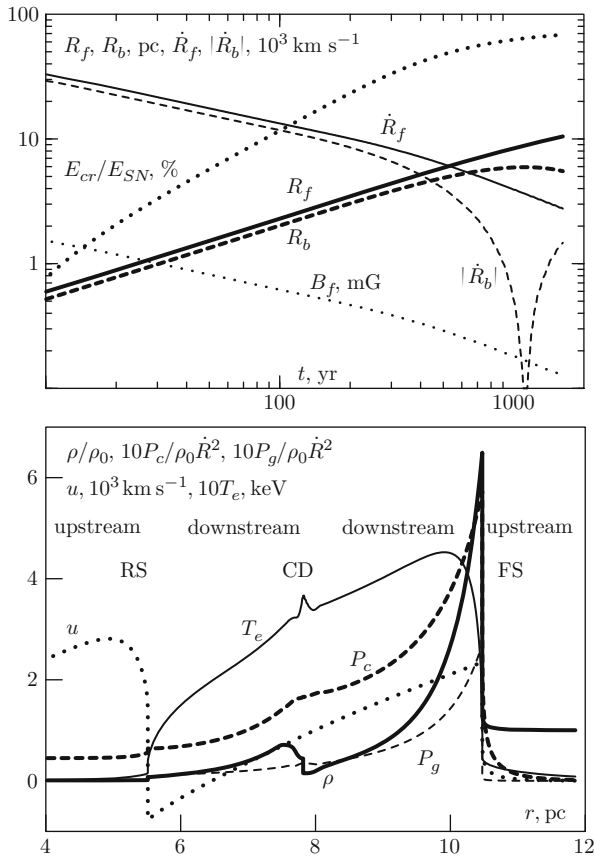
The inner rings seen in the radio and X-ray images of several SNRs, in particular in Cas A (Helder and Vink 2008), are most naturally explained by reverse shocks. Generally, the reverse shock is not treated as an effective accelerator because the magnetic field of the ejecta is expected to be very weak. However, similar to the forward shock, the magnetic field in the reverse shock can be significantly amplified as well. Moreover, the particle acceleration in reverse shock can be quite effective and compete with acceleration of protons and electrons by the forward shock (Zirakashvili and Aharonian 2010). The parameters characterizing the reverse shock can be significantly different compared to the parameters of the forward shock. Generally, the magnetic field and the plasma density in the reverse shock are expected to be very low; this can provide very high efficiency for IC gamma-rays, and, at the same time, dramatic suppression of production of hadronic gamma-rays. And vice versa, because of the higher gas density and stronger magnetic field, the contribution of the hadronic component in the forward shock can dominate over the flux of IC gamma-rays. The detailed numerical studies of particle acceleration and radiation in forward and reverse shocks applied to RX J1713.7-3946 (Zirakashvili and Aharonian 2010) show that the reverse shock can indeed introduce dramatic changes in the spectral and spatial distributions of radiation.

► *Figure 15-8* shows the time and radial dependences (at the present epoch $t = 1,600$ year) of basic parameters characterizing the forward and reverse shocks. They are obtained for the ejecta mass $M_{\text{ej}} = 1.5M_{\odot}$ and the energy explosion $E_{\text{SN}} = 2.7 \cdot 10^{51}$ erg, assuming that the SNR shock propagates through the medium of temperature $T = 10^4$ K with gas density $n_{\text{H}} \approx 0.1 \text{ cm}^{-3}$.

The energy spectra of accelerated protons and electrons are shown in ► *Fig. 15-9* after 100 years of explosion and at the present epoch $t \sim 1,600$ years. While the maximum energy of protons at $t = 100$ years, when the shock speed was $\approx 1.3 \cdot 10^4 \text{ km s}^{-1}$, exceeds 600 TeV, at the present epoch it is significantly smaller, $E_{\text{max}} \sim 150$ MeV. The reason is that the higher energy particles have already left the remnant.

It is interesting to note that at $t = 100$, year the electrons in the inverse shock dominate over the electrons accelerated by the forward shock (apparently because of very strong synchrotron losses in the latter), while the contributions of the forward and reverse shocks to accelerated protons are comparable. At the present epoch, the forward shock contributes more to the acceleration, except for the highest energy particles. Apparently this should have an impact on the energy spectrum and morphology of high-energy radiation. While the contribution of the reverse shock to the hadronic component of radiation is small because of the low-density ambient plasma, it can be comparable or even exceed the contribution from the forward shock as long as this concerns the highest energy tails of the synchrotron X-rays, and especially the IC gamma-rays.

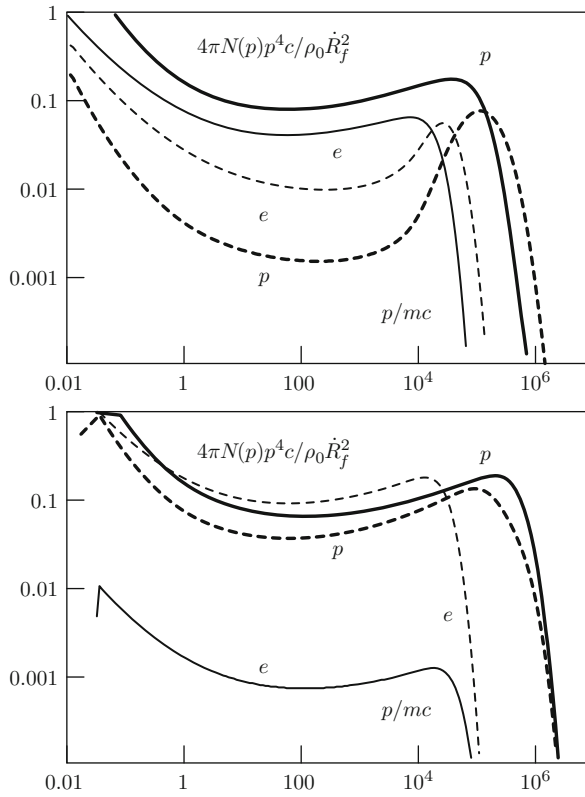
The SEDs of broadband radiation of RX J1713.7-3946 calculated for a model with parameters which allow explanation of gamma-radiation by hadronic interactions of accelerated protons with ambient gas (top) and IC scattering of electrons (bottom panel) are shown in ► *Fig. 15-10*. The hadronic interpretation of gamma-rays requires strong amplification of the magnetic field, and high efficiency of proton acceleration. Given the low density of the ambient gas, $n \approx 0.1 \text{ cm}^{-3}$, it is crucial that at the present epoch almost 70% of the explosion energy of 2.7×10^{51} erg is already transferred to accelerated protons and most of them are still confined in the shell of the remnant (see ► *Fig. 15-9*). The injection efficiency of electrons is adjusted



■ Fig. 15-8

Time and radial dependencies of parameters characterizing the forward and reverse shocks (from Zirakashvili and Aharonian (2010)). *Top panel*: time-dependences of the forward and reverse shock radii, R_f and R_b , and velocities \dot{R}_f and \dot{R}_b , as well as the magnetic field strength downstream of the forward shock B_f , and the ratio of the CR energy to the total energy of the supernova explosion E_{cr}/E_{SN} . *Bottom panel*: radial distributions of the gas density ρ , temperature T_e , pressure P_g , and speed u , as well as the CR pressure P_c in the remnant at the present epoch of evolution $t = 1,620$ years. The contact discontinuity (CD) between the ejecta and the interstellar gas is situated at $r = R_c = 7.8$ pc. The reverse shock (RS) in the ejecta is situated at $r = R_b = 5.5$ pc

in order to reproduce the total flux of nonthermal X-ray emission. The electron injection efficiency at the reverse shock is adjusted to reproduce the observable radio-intensity of the inner ring. Although the reverse shock contributes significantly to the highest energy particles, especially around 100 TeV (see ► Fig. 15-9), the gamma-ray production related to this population of protons is suppressed because of the low density of the ejecta. The contribution of electrons directly accelerated by the reverse shock to synchrotron radiation and IC gamma-rays is more significant.



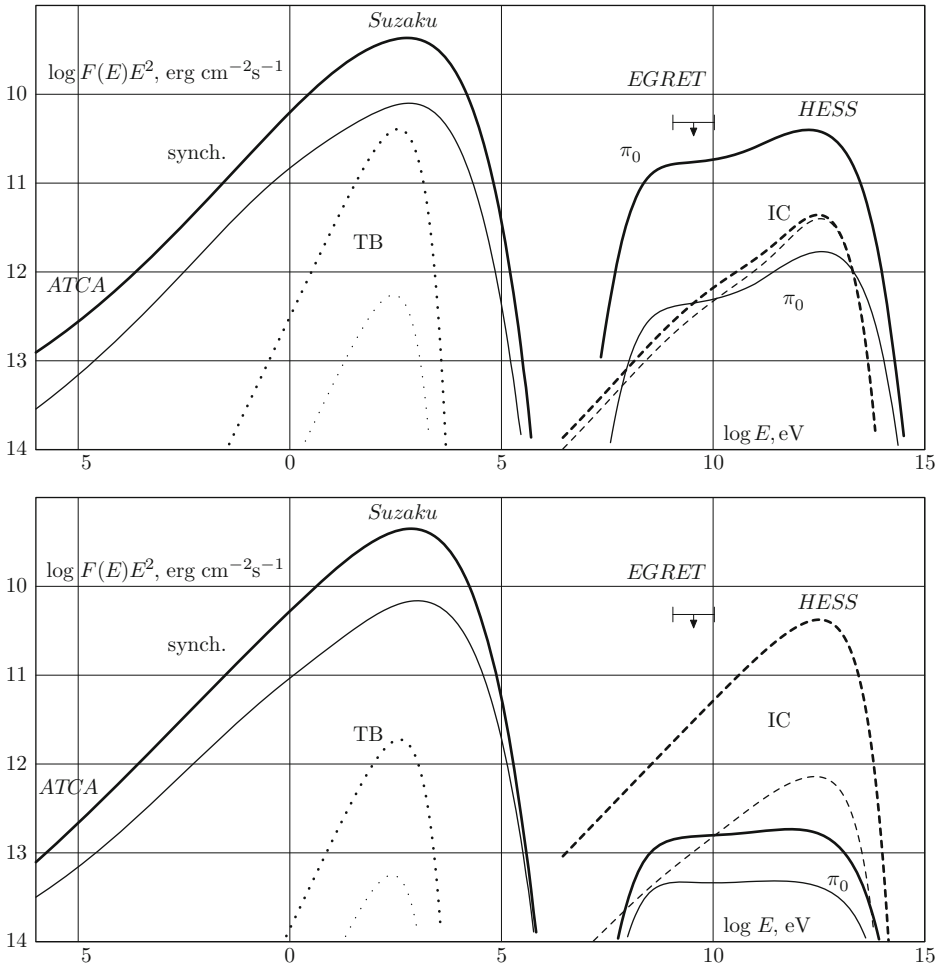
■ Fig. 15-9

The energy distributions of accelerated protons (*thick lines*) and electrons (multiplied to the factor of 5,000, *thin lines*) at $t = 100$ year (*top panel*) and at the epoch $t = 1,620$ year (*bottom panel*). Spectra at both the forward shock (*solid lines*) and the reverse shock (*dashed lines*) are shown

The characteristic features of the hadronic model is the strong amplification of the magnetic field and the huge energy released in very high-energy protons, $W_p \sim 10^{51}$ erg. As mentioned above the model demands also uncomfortably large proton-to electron ratio, $p/e \sim 10^4$.

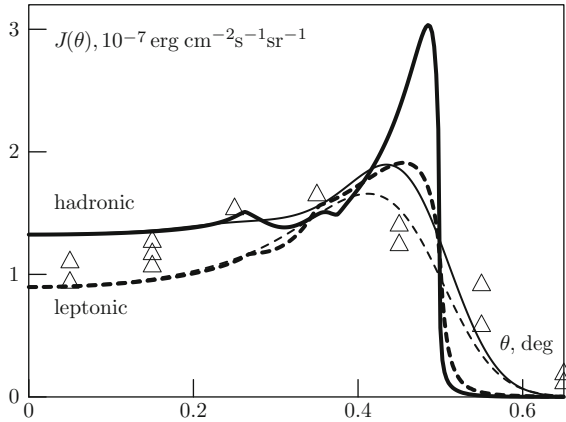
In this regard, the leptonic scenario, which assumes that gamma-rays are produced predominantly due to the inverse Compton scattering of electrons, requires less extreme assumptions. As in the case of hadronic model, both the forward and inverse shocks contribute to the acceleration of electrons and protons (see ● Fig. 15-9). An interesting consequence of this model is the weak thermal emission, and proton/electron ratio close to 100, both giving some preference to the leptonic model versus the hadronic one.

While the observed energy distribution of gamma-rays only marginally can be accommodated within the IC model, the detected broad spatial distribution of TeV radiation agrees quite well with predictions of this model (see ● Fig. 15-11). The low magnetic field, which is a key element of the IC model, allows multi-TeV electrons propagate to large distances, and thus to fill quite a large volume. Because of homogeneous distribution of the target photon fields, the spatial distribution of resulting IC gamma-ray appears quite broad. In contrast, while the



■ Fig. 15-10

Broadband Spectral energy distributions produced within the hadronic and leptonic models (Zirakashvili and Aharonian 2010). The contributions of radiation processes in the forward (*thick lines*) and reverse (*thin lines*) shocks are shown: proton–proton interactions (π^0 ; *solid lines*), inverse Compton scattering (IC; *dashed lines*), synchrotron radiation (synch.; *solid lines*) and thermal bremsstrahlung (TB; *dotted lines*). *Top panel*: Hadronic model. The key model parameters adopted to match the observed fluxes are: $n_H = 0.09 \text{ cm}^{-3}$, $E_{SN} = 2.7 \cdot 10^{51} \text{ erg}$, $M_{ej} = 1.5 M_{\odot}$. The downstream magnetic fields and the speeds of the forward and reverse shocks at the present epoch are: $B_f = 127 \mu\text{G}$ and $B_b = 21 \mu\text{G}$, $V_f = 2,760 \text{ km s}^{-1}$ and $V_b = -1,470 \text{ km s}^{-1}$, respectively. *Bottom panel*: SED produced in the leptonic model. The adopted key model parameters are: $n_H = 0.02 \text{ cm}^{-3}$, $E_{SN} = 1.2 \cdot 10^{51} \text{ erg}$, $M_{ej} = 0.74 M_{\odot}$. The downstream magnetic fields and the speeds of the forward and reverse shocks at the present epoch obtained in calculations are: $B_f = 17 \mu\text{G}$ and $B_b = 31 \mu\text{G}$, $V_f = 3,830 \text{ km s}^{-1}$ and $V_b = -12,200 \text{ km s}^{-1}$, respectively



■ Fig. 15-11

Radial profiles of 1 TeV gamma-rays calculated for the hadronic and electronic scenarios (Zirakashvili and Aharonian 2010) in the uniform medium (*solid*) and for the leptonic scenario with the unmodified forward shock (*dashed*). The profiles smoothed with a Gaussian point spread function with $\sigma = 0.05^\circ$ are also shown (*thin lines*). The *triangles* correspond to the azimuthally averaged TeV gamma-ray radial profile as observed by HESS

hadronic model provides a better spectral fit, it predicts narrower spatial gamma-ray distribution, mainly due to the enhanced emission in the compressed region of the shock, as it is seen in [Fig. 15-11](#). However, because of limited angular resolution of gamma-ray telescopes, it is hard to distinguish between the radial distributions predicted by two models. This is demonstrated in [Fig. 15-11](#) where the radial profiles are smoothed with a typical for the current Cherenkov telescope arrays point spread function of $\sigma = 3$ arcmin. Both smoothed profiles reasonably agree with the angular distribution of TeV gamma-rays. The improvement of angular resolution of future gamma-ray detectors by a factor 1.5–2 should allow decisive conclusions concerning the origin of parent particles based on the gamma-ray morphology.

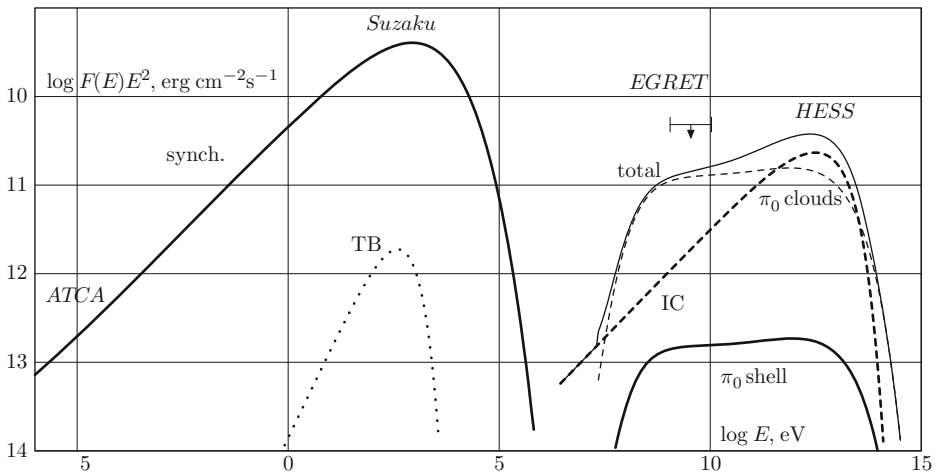
It should be noted however that the reduction of the issue to the question “hadronic or electronic?” could be an oversimplification of the problem. In fact, the formation of spectral and spatial distributions of high-energy radiation in SNRs might proceed in a quite complex way, with significant contribution from both electron and proton populations of accelerated particles. This, in particular, could be the case of RX J1713.7-3946, given the possible containment of shocked dense clouds in the low-density shell as a whole.

Given the estimated explosion energy of RX J1713.7-3946, $E_{SN} < 3 \cdot 10^{51}$ erg, the ejecta mass must be small, $M_{ej} < 2M_{\odot}$, otherwise the forward shock speed would be too small for acceleration of particles well above 10 TeV. Such small ejecta masses correspond to Ib/c or IIb core collapse supernova. Circumstellar medium around these types of supernova is created by the stellar wind of the supernova progenitor. It may be a very low density bubble, $n_H \sim 0.01 \text{ cm}^{-3}$, formed by the stellar wind of a Wolf-Rayet progenitor. However, if the bubble is formed during the Red Supergiant stage of the supernova progenitor or during the interaction of the slow Red Supergiant wind with the fast Wolf-Rayet wind, the gas density could be much higher. It is quite possible that the shock must have swept the progenitor’s material ejected during the stellar evolution, while the interaction with the molecular gas surrounding the remnant only starts.

The exceptions could be the very dense cores of molecular gas (clouds C and D) that are situated inside the forward shock (Fukui 2008). Probably the high gas pressure from the downstream region has driven secondary shocks into the clouds as discussed by Chevalier (1977). The non-thermal X-rays can be produced at these shocks by high-energy electrons from the remnant's shell in the compressed magnetic field of the cloud, while the highest energy gamma-rays from pion decay are mainly produced inside the cloud where the target density is high (Fukui 2008).

If the highest energy protons freely penetrate into the clouds, the corresponding gamma-emission from the pion decay may exceed the gamma-emission from the pion decay of the remnant by a factor that is the ratio of the cloud mass to the mass swept up by the forward shock. Thus, in the scenario with very low gas density of the shell, i.e., when the overall gamma-ray emission is dominated by the IC scattering of electrons, in addition to the leptonic component of radiation one may expect significant contribution of hadronic gamma-rays produced in dense gas condensations.

• Figure 15-12 demonstrates the feasibility of such a “composite” scenario. The ratio of gamma-rays from clouds to the flux from other parts of the shell depends on the ratio of the mass of clouds to the mass of the shell, provided that all particles freely enter the dense clouds. This however could not be the case, especially for the low-energy particles. Because of slow diffusion, the penetration of low-energy particles into the dense cores of these condensations can take longer than the age of the SNR. Correspondingly, the low-energy gamma-ray emission can be suppressed. This effect can offer a reasonable explanation of the hard gamma-ray spectrum as reported recently by the *Fermi* collaboration in the GeV energy region.



■ Fig. 15-12

Broadband emission of RX J1713.7-3946 for the “composite” scenario of gamma-rays with a non-modified forward shock and dense clouds. The principal model parameters are: $t = 1,620$ year, $D = 1.5$ kpc, $n_H = 0.02$ cm $^{-3}$, $E_{SN} = 1.2 \cdot 10^{51}$ erg, $M_{ej} = 0.74 M_{\odot}$, $B_f = 22$ μ G and $B_b = 31$ μ G, $V_f = 3,830$ km s $^{-1}$, $V_b = -1,220$ km s $^{-1}$. The hadronic component of gamma-rays from the remnant's shell is shown by a solid line, and from dense clouds assuming flux enhancement by the factor of 120 is shown by a dashed line

Independent of the ability of different models to describe the spectral and morphological features of gamma-ray emission of RX J1713.7-3946, it is obvious that we deal with a source that effectively accelerates electrons and protons to energy of 100 TeV and beyond. On the other hand, this object has many unique features, and in this regard it could be misleading if we treat this source as a representative of the whole population of young SNRs. Moreover, one should be prepared that the gamma-ray production in this unique object might proceed in a quite peculiar way. In this regard one should mention the interesting idea suggested by Malkov et al. (2005) which can explain in a quite natural way both the low synchrotron radio flux and lack of thermal X-ray emission in RX J1713.7-3946. The standard scenarios of gamma-ray production in SNRs assume that radiation is produced downstream where the densities of both relativistic particles and thermal plasma are higher than in upstream. However, when the shock is expanding into a low-density wind bubble and approaching cold dense material, e.g., the swept-up shell or surrounding molecular clouds, the gamma-radiation is contributed predominantly from upstream. While the energy distribution of accelerated particles downstream is coordinate-independent (in both linear and nonlinear regimes), the particle distribution in the upstream region is coordinate-dependent. Because of the energy-dependence of the diffusion coefficient, the high-energy particles diffuse ahead of low-energy particles, thus a dense material adjacent upstream will “see” relativistic particles (protons and electrons) with low-energy cut-offs. This implies that the effective production of TeV gamma-rays (from p-p interactions) and X-rays (from synchrotron radiation of TeV electrons) will be not accompanied by low-energy gamma-rays and synchrotron radio emission, as well as by thermal X-rays.

7 Cas A

The shell-type supernova remnant Cas A (Cassiopeia A) is one of the best studied nonthermal objects in our Galaxy. Some of its general features are common, to a certain extent, for other young SNRs as well. Yet, Cas A is a rather unique representative of the remnants of supernovae explosions.

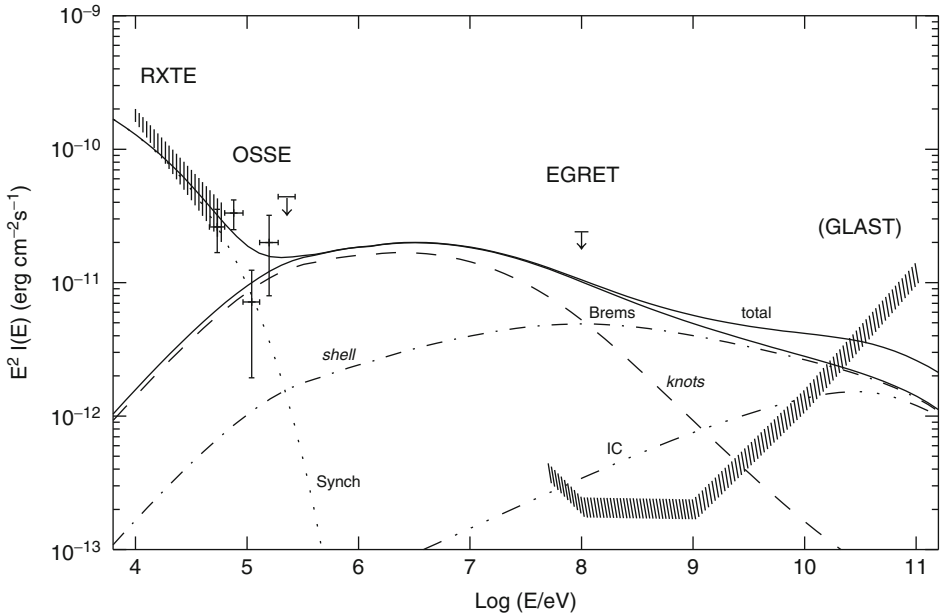
The synchrotron emission of Cas A spans from radio to hard X-rays indicating the presence of relativistic electrons from sub-GeV to multi-TeV energies. For any reasonable assumption on the nebular magnetic field, this object contains enormous energy in the form of relativistic electrons – as large as 3×10^{48} erg (Atoyan et al. 2000a). The rate of accumulation of this energy over the short time period, $t_{\text{acc}} \leq t_{\text{age}} \approx 330$ years, is even more striking, $\dot{W}_e = W_e/t_{\text{acc}} \approx 3 \times 10^{38}$ erg s⁻¹. It is larger, by at least an order of magnitude, than the electron production rate in any other supernova remnant. On the other hand, the content of protons in this object could be relatively modest. As discussed below, the amount of relativistic protons and nuclei is constrained by gamma-ray fluxes detected in the GeV and TeV energy bands, $W_p \leq 3 \times 10^{49} (n/10\text{cm}^{-3})^{-1}$ erg (Abdo et al. 2010b). This constitutes less than 2% of the total explosion energy, if gamma-rays are indeed produced in the reverse shock where the plasma is quite dense, $n \geq 10$ cm⁻³. In this case, the ratio of relativistic protons to electrons in Cas A is less than 10, i.e., an order of magnitude below the level observed in cosmic rays. This is in sharp contrast to the hadronic models of gamma-radiation of other SNRs, e.g., SN 1006 and RX J1713.7-3946, which require $p/e \geq 10^3$. However, if the reported GeV and TeV gamma-ray fluxes are produced at hadronic interactions in the forward shock, which propagates through low-density circumstellar medium, the total energy in accelerated protons can exceed 10^{50} erg.

Another unique feature of this object is the reported X-ray (Renaud et al. 2006) and gamma-ray (Iyudin et al. 1994) emission lines associated with ^{44}Ti . This radiation component provides direct information about the ejected mass of radioactive titanium-44, $M_{44\text{Ti}} \approx 2 \times 10^{-4} M_{\odot}$. In this regard Cas A seems to be a rather peculiar object compared to other young SNR from which so far no characteristic emission of ^{44}Ti has been unambiguously detected.² It has been recently suggested (Zirakashvili and Aharonian 2011) that the enormous content of relativistic electrons in Cas A could have an interesting link to the ejection of large amount of radioactive material, first of all ^{44}Ti and ^{56}Ni . The decay products of these nuclei provide a vast pool of suprathermal positrons and electrons which can be further accelerated to multi-TeV energies by reverse and forward shocks.

The GeV and TeV electrons which are responsible for the broadband synchrotron radiation of Cas A, inevitably produce also gamma-rays – through bremsstrahlung and inverse Compton scattering. The calculations of gamma-ray fluxes based on the radio data are straightforward, but strongly depend on distributions of the gas and magnetic field in the nebula. The bulk of radiation of Cas A of both thermal and nonthermal origin comes from the shell enclosed between two spheres with angular radii of 100 and 150 arcsec (corresponding to spatial radii $R = 1.7$ pc and $R = 2.5$ pc for the distance to the source $d = 3.4$ kpc). A major fraction of nonthermal energy originates not only in the shell, through the diffusive shock acceleration, but perhaps also in the numerous hot spots which appear to be fast-moving knots or compact, steep-spectrum radio structures (see, e.g., Cowsik and Sarkar (1984)). The radio structures are of special interest because it is likely that their bright radio emission is not just a result of enhanced magnetic field, but is (also) caused by the local enhancement of relativistic electrons. What concerns the steep radio spectra, the energy distributions of parent electrons in these compact structures could become significantly steeper (compared to the acceleration spectrum) because of the *energy-dependent* escape of electrons into the surrounding diffuse plateau region. The discovery of variations of nonthermal emission of X-ray filaments and knots seen in *Chandra* data on year timescales (Uchiyama and Aharonian 2008) provide further support for the hypothesis that these hot spots can be sites of particle acceleration (Atoyan et al. 2000a). Indeed if the variations of the X-ray flux would be caused by the change of the magnetic field, we should see correlated similar (or stronger) variations also in the radio band as well. However, the archives of the VLA data at 5 GHz frequency do not show significant brightness changes on timescales less than 30 years, therefore it seems more natural to relate the observed X-ray variability to the increase (via electron acceleration) or decrease (via radiative cooling or escape) of multi-TeV electrons.

The spectral and temporal evolution of synchrotron radiation of Cas A within the two-zone model, which distinguishes between compact bright steep-spectrum radio knots (zone 1) and the diffuse “plateau” (zone 2), have been used by Atoyan et al. (2000b) to predict gamma-ray fluxes produced by electrons responsible also for the broadband synchrotron emission (see [Fig. 15-13](#)). The total gamma-ray flux is formed from contributions of bremsstrahlung and inverse Compton components produced in both zones. Below 10 GeV, the gamma-ray flux is strongly dominated by electron bremsstrahlung. Because of the steep decline of the energy distribution of radio electrons in the compact radio structures, the intensity of the gamma-ray flux at $E \sim 1$ GeV is dominated by the flat-spectrum bremsstrahlung of the plateau region.

²Recently, the detection of 4.1 keV line emission has been reported from the youngest galactic supernova remnant G1.9+0.3 (Borkowski et al. 2010). The likely interpretation of this line is its association to ^{44}Sc which is produced from ^{44}Ti via electron capture.



■ Fig. 15-13

The fluxes of synchrotron (dotted line), inverse Compton (3-dot-dashed line), and bremsstrahlung (solid line) radiations calculated in the framework of the two-zone model (Atoyan et al. 2000b). The broadband overall spectral energy distribution consisting of these three components of radiation is shown by the curve marked as “total.” The bremsstrahlung fluxes produced in zone 1 and zone 2 are shown separately by the dashed and dot-dashed curves, respectively. The hatched region shows the flux sensitivity anticipated for *Fermi* LAT (that time GLAST). The X-ray/soft gamma-ray fluxes measured by RXTE and OSSE detectors, as well as the flux upper limit from EGRET, are also shown

It is also important that the contribution of other gamma-ray production mechanisms at this energy is not yet significant. This makes the prediction of gamma-ray fluxes in the GeV band quite robust. Remarkably, both the level of the flux and the energy spectrum of gamma-rays recently reported by the *Fermi* collaboration (Abdo et al. 2010b) match well with the theoretical predictions presented in ▶ Fig. 15-13.

The situation in the high-energy domain is more complicated. Since the synchrotron spectrum extends to hard X-rays, we should expect extension of the gamma-ray spectrum to TeV energies. Along with bremsstrahlung, the principal production mechanism at these energies is the inverse Compton scattering of electrons in the field of thermal dust emission with $T \simeq 100$ K. The extrapolation of gamma-ray fluxes shown in ▶ Fig. 15-13 toward TeV energies is significantly below the detection threshold of ground-based detectors. While it is difficult to increase the contribution of bremsstrahlung, in principle there is a room to enhance the IC contribution. The fluxes of IC gamma-rays are very sensitive to the mean magnetic field in the shell. In particular, in the VHE regime, $F_{IC} \sim B^{-(5+\alpha_2)/2}$, where $\alpha_2 \geq 3$ is the spectral index of TeV electrons in the shell. Thus, the flux of IC gamma-rays at TeV energies can be boosted assuming weaker magnetic field. On the other hand, the magnetic field cannot be significantly less than 0.3 mG, otherwise the contribution of bremsstrahlung would lead to overproduction of X-ray

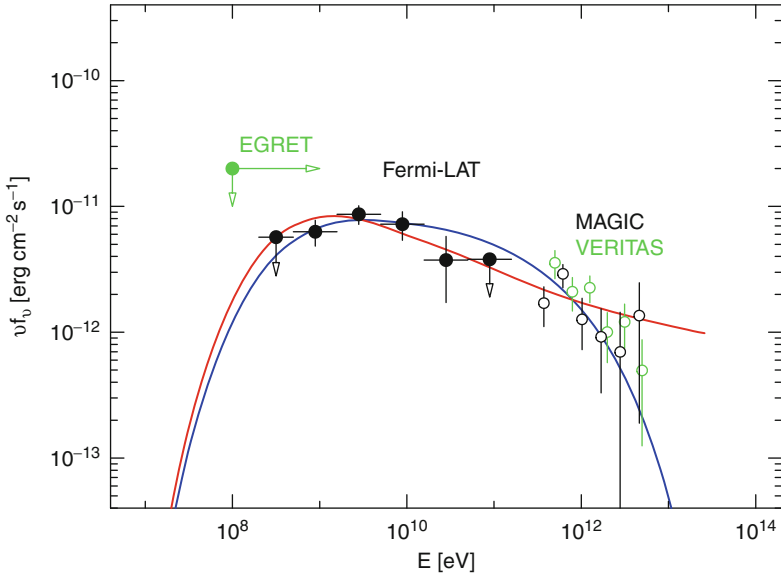
and low-energy (GeV) gamma-rays. Therefore a quite robust conclusion from analysis of spectral and morphological features of synchrotron emission is that one cannot expect detectable TeV gamma-ray emission of IC origin, unless we invoke more zones, i.e., regions with low magnetic fields but yet with adequate conditions for effective acceleration of electrons to multi-TeV energies. Although at first glance this sounds a rather superficial assumption, the regions with very low magnetic fields in SNRs cannot be a priori excluded. Moreover, in the case of Cas A this can be realized in a quite natural way, namely, through the operation of the reverse shock.

Although the initial magnetic field in reverse shocks is expected to be very weak (due to the expansion of the ejecta), later the field can be amplified, like in forward shocks, by different instabilities introduced, e.g., by accelerated particles. Another effect related to accelerated particles is the modification of the reverse shock by relativistic protons and nuclei (as discussed above for RX J1713.7-3946). Remarkably, in Cas A the reverse shock can be modified not only by nucleonic component of accelerated particles, but also by electrons (Zirakashvili et al. 2011). The shock modification is a principal issue, especially in Cas A, for explanation of steep spectra of electrons at low energies as it follows from radio observations. Note, however, that in Cas A one should distinguish between several populations of radio-emitting electrons. In the case of compact radio knots, the very steep radio spectra can be explained also by the energy-dependent escape of particles from these compact structures (Atoyan et al. 2000a).

The relatively steep spectrum of synchrotron radio emission of the forward shock can be of different origin. In particular, it could be related to the azimuthal magnetic field of the medium where the forward shock propagates. Since in this case the mean magnetic field is parallel to the shock surface, the propagation of charged particles is described by the so-called compound diffusion which results in particle spectra close to $E^{-2.5}$ (Kirk et al. 1996). While such a spectrum describes quite well the radio emission of the forward shock, it is too steep compared to the gamma-ray spectrum reported at GeV energies (Abdo et al. 2010b).

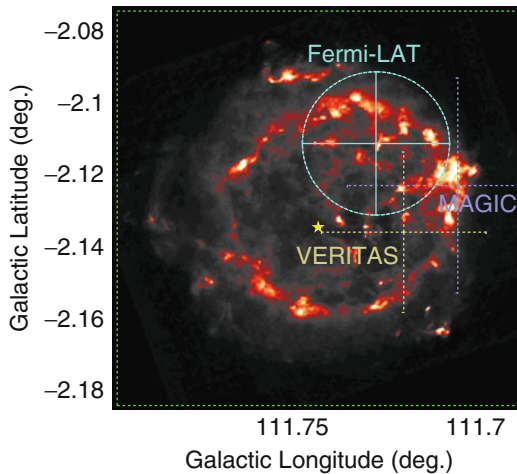
The flux measured by *Fermi* in the energy interval between 0.3 and 30 GeV is flat with a slope close to $\Gamma = 2.0-2.2$, in contrast to the significantly steeper spectrum reported at TeV energies (see [Fig. 15-14](#)). The VHE gamma-ray signal from Cas A has been discovered in the 1–10 TeV energy interval by the *HEGRA* collaboration (Aharonian et al. 2010). This result later has been confirmed by the *MAGIC* (Albert et al. 2010) and *VERITAS* (Acciari et al. 2010) collaborations which extended the spectral measurements to sub-TeV energies. The fluxes reported by three groups are in good agreement and show relatively steep spectrum with a photon index $\Gamma = 2.4-2.6$. In particular, the spectrum reported by the *VERITAS* collaboration is compatible with power law, $dN/dE \propto E^{-\Gamma}$ with photon index $\Gamma = 2.6$ and total flux above 1 TeV $F_{\gamma}(\geq 1\text{TeV}) = 7.7 \times 10^{-13} \text{ ph cm}^{-2} \text{ s}^{-1}$.

Both the reverse and forward shocks can contribute significantly to the reported GeV and TeV gamma-ray fluxes. Unfortunately the angular resolution of gamma-ray detectors in both energy bands is larger than the angular size of Cas A. While the point spread functions of *Fermi* and Cherenkov telescope arrays are sufficient to be convinced that the positions of GeV and TeV sources are spatially coincident with the radio image of the remnant (see [Fig. 15-15](#)), the gamma-ray data do not allow us to separate the contributions of different regions to the overall gamma-ray flux. This introduces significant arbitrariness in the interpretation of the origin of gamma-ray emission. While the contribution of the electron bremsstrahlung from the reverse shock to the GeV gamma-ray flux can be close to 100%, the explanation of TeV gamma-ray fluxes requires an additional radiation component(s) related to the inverse Compton scattering and/or π^0 -decay gamma-rays. Because of the strong magnetic field of order of 1 mG, the efficiency of production of IC gamma-rays in the forward shock is expected to be small.



■ Fig. 15-14

The differential energy spectra of gamma-rays reported by the *Fermi*, *MAGIC*, and *VERITAS* collaborations (from Abdo et al. (2010b)). Two curves correspond to calculations of theoretical gamma-ray spectra from p–p interactions assuming for protons (1) a single power law distribution with an index $\alpha = 2.3$ (red line) and (2) a power-law distribution with $\alpha = 2.1$ an exponential cutoff at $E_0 = 10$ TeV



■ Fig. 15-15

The VLA radio map of Cas A with indication of positions of the detected GeV (Abdo et al. 2010b) and TeV (Abdo et al. 2010b; Acciari et al. 2010) gamma-ray signals (error bars including both the statistical and systematic errors)

In this regard the reverse shock “offers” better conditions for realization of the IC scenario. And vice versa, despite the existence of high-density plasma in the reverse shock, the conditions for gamma-ray production via p–p interactions are settled better in the forward shock (Zirakashvili et al. 2011).

Independent of the issue related to the location(s) of gamma-ray production, a hard proton spectrum with an exponential cutoff around 10 TeV can nicely explain the overall GeV–TeV gamma-ray spectrum, as it is demonstrated in [Fig. 15-14](#). Note that the total energetics in protons GeV and TeV which is required to explain the absolute gamma-ray fluxes, $W_p \approx 3 \times 10^{49} (n/1 \text{ cm}^{-3})^{-1} \text{ erg}$, depends on the density of the ambient medium. If gamma-rays are produced in the reverse shock with gas density exceeding 10 cm^{-3} , the total energy in accelerated protons and nuclei constitutes only 2% of the explosion energy.

On the other hand, if gamma-rays have hadronic origin, and they are produced in the forward shock propagating through the circumstellar medium of density $\sim 1 \text{ cm}^{-3}$, the total energy in protons is increased by an order of magnitude, i.e., can be as large as 20% of the explosion energy. In this case, the proton-to-electron ratio also becomes quite close to the “nominal” value of 100. This seems a quite attractive option, at least in the context of the SNR paradigm of the origin of cosmic rays. Even so, the required cutoff in the proton spectrum at 10 TeV is quite disappointing; there is no evidence that Cas A operates as a PeVatron. This is true also for other young SNRs. Whether this is connected with the escape of highest energy particles accelerated at the early epochs or there are some other reasons, we do not yet know.

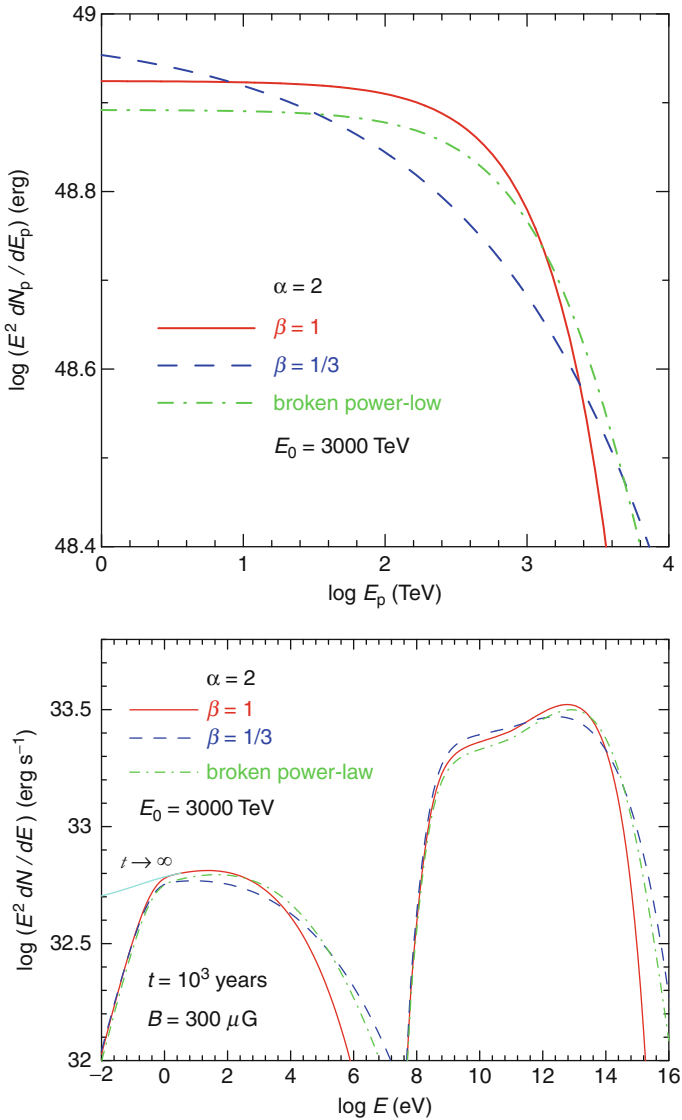
It is interesting to note that from the point of view of SNR paradigm of galactic cosmic rays, the interpretation of current gamma-ray data within the leptonic (but not hadronic!) models is a more preferable option. This, at first glance paradoxical conclusion implies that protons are, in fact, accelerated out to 1 PeV, but we do not see the related π^0 decay gamma-rays because of very low density ambient gas density.

Meanwhile, without answers to these question, we cannot be sure that SNRs are the major contributors to the galactic cosmic rays up to the “knee” around 1 PeV. The “hunt” for galactic PeVatrons continues.

8 Searching for Galactic Proton PeVatrons

The most straightforward search for galactic PeVatrons can be conducted by ground-based gamma-ray detectors designed for operation in the energy regime between 10 and 100 TeV. Generally, the identification of the hadronic component against the competing IC radiation is a difficult task. Fortunately, at gamma-ray energies above 10 TeV, the IC background is gradually faded out. The reason is that in SNR shocks with relatively low acceleration rate, the synchrotron losses prevent acceleration of electrons to energies beyond 100 TeV. In other words, SNRs cannot act as *electron PeVatrons*.³ Also at such high energies the contribution of the IC component is suppressed because of the Klein–Nishina effect. Thus, the hadronic origin of such energetic gamma-ray signals from SNRs hardly could be debated. [Fig. 15-16](#) shows the X-ray and gamma-ray luminosities of hadronic origin from a 1,000-year-old PeVatron calculated for

³Note that in pulsar wind nebulae the acceleration rate of electrons exceeds, for the given magnetic field, the corresponding rate in SNR shocks of speed v_{sh} by a factor of $(c/v_{\text{sh}})^2 \sim 10^3 - 10^4$; therefore in these objects electrons can be accelerated to PeV energies. The *HESS* observations of several plerions with hard gamma-ray spectra extending to tens of TeV without indication of a spectral cutoff is indeed a strong evidence that pulsar wind nebulae act as extremely effective electron accelerators.



■ Fig. 15-16

The broadband radiation of a PeVatron initiated by interactions of protons with the ambient gas. *Top panel:* three different spectral distributions of protons: “power-law with exponential cutoff” in the general form $E^{-\alpha} \exp(-E/E_0)^\beta$ with $\alpha = 2$, $E_0 = 3$ PeV, and $\beta = 1$ (solid curve), $\beta = 1/3$ (dashed curve), and “broken power-law” when the spectral index is changed at $E = 1$ PeV from $\alpha = 2$ to $\alpha = 3$. *Top panel:* the luminosities of broadband emission initiated by interactions of protons with the ambient gas, including the decays of neutral pions to gamma-rays and synchrotron radiation of secondary electrons from the decays of charged pions. The ambient density is assumed $n = 1 \text{ cm}^{-3}$ and the age of the source $t = 10^3$ years

three different distributions of accelerated protons (shown in the same figure). Both radiation components are initiated by interactions of accelerated protons assuming the ambient gas density $n = 1 \text{ cm}^{-3}$ and magnetic field $B = 300 \mu\text{G}$. While gamma-rays arise directly from the decay of π^0 -mesons, X-rays are a result of synchrotron radiation of secondary electrons, the products of π^\pm -decays. The lifetime of electrons producing X-rays, $t_{\text{synch}} \simeq 1.5 B_{\text{mG}}^{-3/2} (E_X/1 \text{ keV})^{-1/2} \text{ year}$, in a magnetic field exceeding $100 \mu\text{G}$ is very short (≤ 50 year) compared to the age of the source. Therefore the synchrotron X-radiation actually could be considered as an unavoidable “prompt” radiation component of hadronic interactions, emitted simultaneously with gamma-rays from the π^0 -decays and neutrinos from π^\pm -decays.

The X-ray and gamma-ray fluxes depend on the total amount of high-energy protons currently accumulated in the source, and on the density of the ambient matter. Although approximately the same fraction of energy of the parent protons is transferred to secondary electrons and gamma-rays, because the energy of relatively low-energy (sub-TeV) electrons is not radiated away effectively, the direct (π^0 -decay) gamma-ray luminosity exceeds the synchrotron luminosity. The L_X/L_γ ratio depends on the proton spectrum as well as on the particle injection history. For the energy distributions of protons shown in [Fig. 15-16](#), the energy release in the X-ray channel is about one fifth of the gamma-ray luminosity.

The spectrum of π^0 -decay gamma-rays in the corresponding cutoff region at $\sim (1/10)E_0$ approximately repeats the shape of the proton spectrum around E_0 . Thus the search for gamma-rays of energy $E \geq 30 \text{ TeV}$ would lead to the discovery and identification of galactic sources responsible for the CR spectrum up to the knee. Moreover, accurate spectroscopic measurements at highest energies would provide extremely important information about the shape of the source (acceleration) spectra of protons. This information is crucial for identification of acceleration mechanisms in SNRs, as well as for understanding the role of different processes (e.g., acceleration versus propagation) responsible for the formation of the knee in the CR spectrum.

The X-ray and gamma-ray luminosities in [Fig. 15-16](#) are calculated for a proton accelerator operating for 10^3 years with a constant rate $L_p = 10^{39} \text{ erg s}^{-1}$; the total energy in protons is $W_p = L_p \cdot T \simeq 3 \times 10^{49} \text{ erg}$. Thus in order to estimate the X-ray and gamma-ray energy fluxes (in units of $\text{erg cm}^{-2} \text{ s}^{-1}$) one should multiply the luminosities in [Fig. 15-16](#) to the factor of $K \approx 10^{-44} (W_p/3 \times 10^{49} \text{ erg})(n/1 \text{ cm}^{-3})(d/1 \text{ kpc})^{-2}$. One can see that for a realistic combination of the product $nW_p \sim 10^{49} - 10^{50} \text{ erg cm}^{-3}$, the future multi-TeV Cherenkov telescope arrays with sensitivities as good as $10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1}$ should be able to probe the ultrahigh-energy radiation of such PeVatrons up to distances of 10 kpc. However, because of very small photon fluxes, it will be quite difficult to perform detailed spectroscopic measurements around the cutoff energy, unless the detection area is dramatically increased to $A_{\text{eff}} \geq 10 \text{ km}^2$ (two orders of magnitude larger than the detection areas of the current Cherenkov telescope arrays). Note that the extension of the proton spectrum to 1 PeV is an important condition for effective production of neutrinos in the 10–100 TeV range which is the most optimal energy interval for TeV neutrino detectors like IceCube or KM3NeT. However, the sensitivity of the neutrino detectors is quite limited, and even the extremely powerful objects in our Galaxy can be only marginally detected by these instruments.

In such circumstances, the search for the PeVatrons via hard synchrotron radiation of secondary (π^\pm -decay) electrons is an alternative, and perhaps even more powerful tool, given the superior potential of X-ray detectors. In particular, *Chandra* and *XMM-Newton* have sufficient sensitivity to perform such studies. However, the effective energy range of these detectors is limited by 10 keV which is not optimal for detection of synchrotron radiation of secondary

electrons, given that a major challenge of this method is the identification of the “hadronic” origin of X-rays at the presence of other X-ray components, in particular its separation from the synchrotron radiation of directly accelerated electrons. These two components can be separated if the magnetic field in the SNR exceeds $100 \mu\text{G}$ and the proton spectrum extends to 1 PeV . These two conditions are, in fact, connected since the acceleration of protons in SNRs to PeV energies is possible only at the presence of large magnetic fields. The second key condition for operation of SNRs as PeVatrons is the diffusion in the *Bohm limit*. In this case, the proton cutoff energy is proportional to the strength of the magnetic field. Thus, the corresponding energy in the spectrum of secondary synchrotron radiation is $h\nu \propto BE_0^2 \propto B^3$. On the other hand, quite remarkably, the position of the cutoff of synchrotron radiation of directly accelerated electrons does not depend on the magnetic field and typically appears in the soft X-ray domain, $h\nu \leq 1 \text{ keV}$ (see, e.g., Aharonian and Atoyan (1999)). Thus, if the spectrum of synchrotron radiation of secondary electrons extends well beyond 10 keV (which surely is the case of proton PeVatrons – see [Fig. 15-16](#)), the problem of the background due to the synchrotron radiation of directly accelerated electrons is dramatically reduced.

The spectrum of the secondary synchrotron radiation in the cutoff region is much smoother and broader than the corresponding cutoff in the gamma-ray spectrum. For example, if the primary protons have a power-law energy distribution with an exponential cutoff written in the general form $(dN/dE)_p \propto \exp[-(E/E_0)^{\beta_p}]$, the distributions of secondary gamma-rays and electrons in the cutoff region is smoother, $(dN/dE)_{\gamma/e} \propto \exp[-(E/E_{0,\gamma/e})^{\beta_{\gamma/e}}]$, with $\beta_{\gamma/e} \approx 0.5\beta_p$ (Kelner et al. 2006). The synchrotron radiation in the corresponding cutoff region behaves as $\exp[-(\varepsilon/\varepsilon_0)^{\beta_s}]$, with $\beta_s = \beta_e/(2 + \beta_e)$ (Zirakashvili and Aharonian 2007). Generally, in the Bohm diffusion regime, $\beta_p = 1$, thus in the cutoff region the distribution of gamma-rays is proportional to $\exp[-(E_\gamma/E_{\gamma/e})^{1/2}]$, while the spectrum of the secondary synchrotron radiation behaves as $\exp[-(\varepsilon/\varepsilon_0)^{1/5}]$. This important feature, which is seen in [Fig. 15-16](#), should allow comprehensive studies of the proton spectra in the cutoff region via X-rays – the third-generation products of p-p interactions. The search for such a component in the nonthermal X-ray spectra of young SNRs is an exciting challenge. If proton PeVatrons of *SNR origin* do exist in our Galaxy, then they should emit hard X-ray emission, and it is quite possible that the galactic proton PeVatrons will be discovered first in X-rays rather than in ultrahigh-energy gamma-rays and neutrinos.

9 Gamma-Ray “Echos” from Nearby Molecular Clouds

The gamma-ray emission from three famous representatives of young SNRs can be interpreted as a result of interactions of shock-accelerated protons with the ambient gas. While quite encouraging, this cannot be considered as a proof of the major contribution of SNRs to the galactic cosmic rays. The problem is that the competing leptonic processes can explain the gamma-ray data as well by IC radiation of directly accelerated electrons. Moreover, within the hadronic models, the required “early” (below 10 TeV) cutoffs in the gamma-ray spectra imply lack of PeV protons in the remnants. A natural reason for the deficit of these most energetic particles could be their leakage from the shell. Actually, the acceleration and confinement of highest energy particles in the remnants can last less than $1,000$ (or, in some cases, even 100) years after the explosion, so one should be lucky to catch up the gamma-ray-emitting SNRs, especially at energies above 10 TeV .

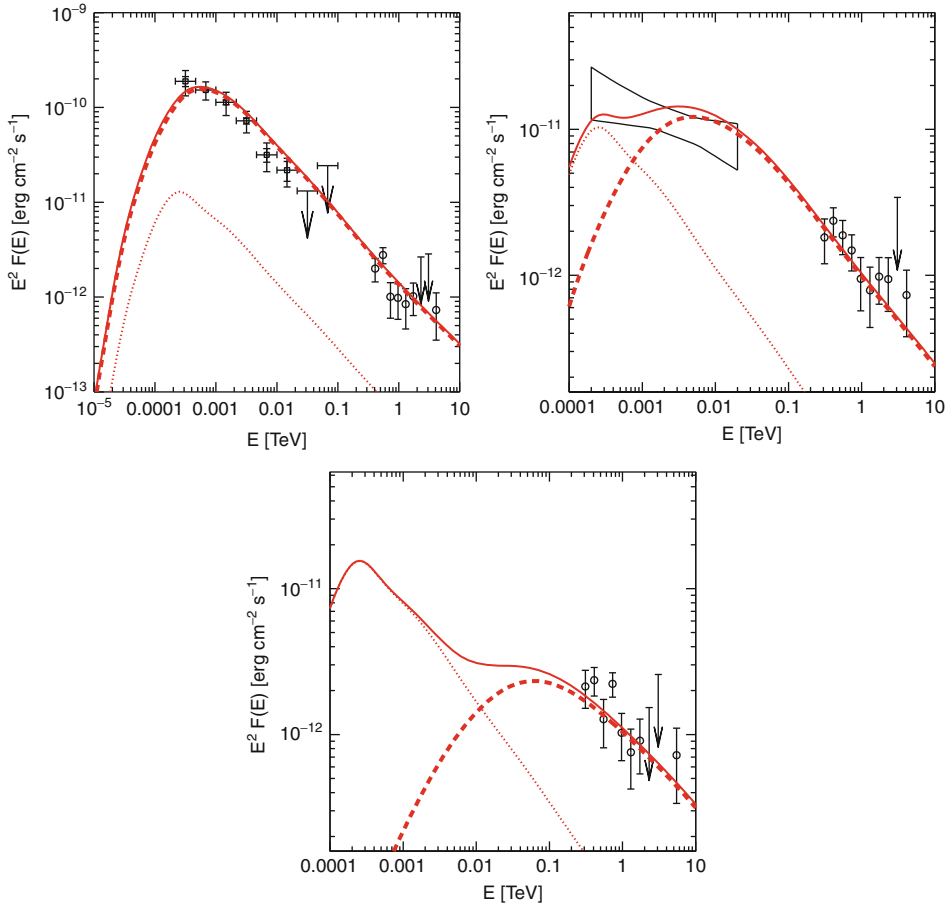
On the other hand, the gamma-ray “echos” of the faded-out accelerators still can be detected, even 10,000 years after the Sedov phase when all particles already have left the remnant. Indeed, the protons, after they escape the SNR shell, diffuse in the interstellar medium and interact with the surrounding atomic and molecular gas. Before being fully diffused away by the interstellar magnetic fields and integrated into the “sea” of galactic cosmic rays, these particles produce gamma-rays the spectrum of which can significantly differ both from the radiation of the SNR itself and from the diffuse emission contributed by the background cosmic ray protons and electrons. The massive molecular clouds (MCs) close to SNR provide dense targets for hadronic interactions, and thus dramatically increase the chances for the secondary gamma-rays to be detected. For typical parameters of a SNR at a distance of 1 kpc, a molecular cloud of mass $10^4 M_{\odot}$ can emit TeV gamma-rays at a detectable level if it is located within a few 100 pc from the SNR (Gabici and Aharonian 2007).

The presence of MCs close to SNRs could be accidental, but in general there is a deep link between SNRs and MCs, especially in the star-forming regions (Montmerle 1979; Paul et al. 1976). Depending on the location of massive clouds, the time of particle injection into the interstellar medium, as well as on the diffusion coefficient, we might expect a broad variety of energy distributions of gamma-rays – from very hard spectra (much harder than the spectrum of the SNR itself) to very steep ones. Correspondingly, the ratio of GeV to TeV gamma-ray fluxes can significantly vary from site to site. Thus the fluxes detected by the *Fermi* and *AGILE* satellites can correlate or anticorrelate with fluxes detected at TeV energies by ground-based detectors. The recent reports from these collaborations support such a picture. For example, if some clouds located close to the mid-age (approximately 10^4 years old) supernova remnant W28 have been reported as gamma-ray emitters both in TeV (Aharonian et al. 2008b) and GeV (Abdo et al. 2010a; Giuliani et al. 2010) bands, in the case of another SNR/MC interacting system, IC 443, the GeV (Tavani et al. 2010) and TeV (Acciari et al. 2009; Albert et al. 2007) gamma-ray images are shifted from each other.

◆ *Figure 15-17* demonstrates that GeV and TeV gamma-ray sources detected around the SNR W28, HESS J1801-233, and HESS J1800-240 A and B can be quite naturally explained by the emission of cosmic rays that escaped several hundred years after the supernova explosion. The gamma-ray images of these sources correlate with CO maps. The masses of the clouds estimated from CO measurements result in five, six, and four times $10^4 M_{\odot}$, respectively; their projected distances from the center of the SNR are 12, 65, and 32 pc, respectively (Aharonian et al. 2008b). Remarkably, for these parameters, and the age of the SNR $\approx 10^4$ years, good fits to the broadband gamma-ray spectra and the absolute gamma-ray fluxes are achieved by variation of only two model parameters: E^{-2} type injection spectrum with total energy $W = 1.2 \times 10^{50}$ erg, and $D(E) \propto E^{0.5}$ type diffusion coefficient with absolute value at 1 GeV $D_0 = 6 \times 10^{26} \text{ cm}^2 \text{ s}^{-1}$. This is by a factor of 20 smaller than the diffusion coefficient in the galactic plane which seems quite reasonable given the much higher turbulence of plasma in the vicinity of W28. Torres et al. (2010) arrived at a similar conclusion concerning the slow diffusion of cosmic rays in the vicinity of SNRs by analyzing the *Fermi*, *AGILE*, *MAGIC*, and *VERITAS* data from IC 443.

One should note that the role of giant molecular clouds is not limited by SNR studies. These massive objects are intimately connected with the star formation regions that are strongly believed to be most probable sites of cosmic ray production (with or without SNRs) in our Galaxy. They serve also as unique “barometers” for measurements of the energy density of cosmic rays in remote parts of the Galaxy (Casanova et al. 2010).

It is generally believed that the local CR flux directly measured at the Earth, gives a correct estimate for the level of the “sea” of galactic cosmic rays. However, strictly speaking, this is an



■ Fig. 15-17

Broadband fit to the gamma-ray emission detected by FERMI and HESS from the sources HESS J1801-233, HESS J1800-240 A and B (*top to bottom*). No GeV emission has been detected from HESS J1800-240 A. *Dashed lines* represent the contribution to the gamma-ray emission from CRs that escaped W28, *dotted lines* show the contribution from the CR galactic background, and *solid lines* represent the total emission (from Gabici et al. (2010))

ad hoc assumption; it is not obvious that local cosmic rays can be taken as being representative of the whole galactic population of relativistic particles. In fact, one cannot exclude that the flux of local cosmic rays could be dominated by a single or few local sources, especially given the fact that the Solar system is located in a rather extraordinary region – inside active star formation complexes which constitute the so-called Gould Belt. The recent *anomalies* discovered in cosmic rays, such as very high content of positrons in the leptonic component of cosmic rays (Adriani et al. 2011a), or the significant differences between energy spectra of protons and alpha particles (Adriani et al. 2011b), tell us that the generally adopted picture of homogeneously distributed galactic cosmic rays contributed by a single class of accelerators (SNRs) could be

an oversimplification. In fact, we may deal with a diverse variety of cosmic ray “factories.” The fortunate location of giant molecular clouds in the vicinity of these accelerators provides us with a unique tool to probe cosmic rays in the environments harboring these mysterious objects.

10 Summary

Despite the intensive efforts of recent years, coordinated by several astronomical communities, we do not have yet a firm evidence of acceleration of nucleonic components of cosmic rays in supernova remnants. The problem here is not related to the question of ability of hadronic models to explain the reported gamma-ray fluxes. In fact, the hadronic models can explain, although with certain caveats, the general features of several gamma-ray-emitting SNRs. The problem is mainly caused by the ability of leptonic models to explain (again, with certain caveats) the same data. The observations of these objects with the next-generation gamma-ray detectors like the Cherenkov Telescope Array (CTA) with significantly improved performance (sensitivity, angular resolution, broader energy coverage) compared to the current instruments should allow us, hopefully, to get clear answers concerning the origin of gamma-radiation of SNRs. A breakthrough in this regard is expected in the so-far unexplored energy region above 30 TeV where the contribution of the IC radiation of directly accelerated electrons is expected to fade out. These energetic gamma-rays contain direct and unambiguous information about protons accelerated to 1 PeV. The detection and identification of the galactic cosmic *PeVatrons* perhaps can be considered as the highest priority objective, at least in the context of the origin of galactic cosmic rays. Together with ultrahigh-energy gamma-rays, the future hard X-ray imaging detectors like *NuStar* and ASTRO-H should be able to conduct an effective search for cosmic *PeVatrons* through the synchrotron radiation of secondary electrons from objects located at distances up to 10 kpc. The coverage of the GeV energy domain by the *Fermi* space telescope is very important, especially for studies of gamma-ray environments which harbor the mysterious cosmic accelerators.

Cross-References

➤ [Mass Distribution and Rotation Curve in the Galaxy](#)

References

- | | |
|--|--|
| Abdo, A. A. et al. (Fermi collaboration) 2010a, <i>ApJ</i> , 718, 348 | Acciari, V. A. et al. (VERITAS collaboration) 2009, <i>ApJ</i> , 698, L133 |
| Abdo, A. A. et al. (Fermi collaboration) 2010b, <i>ApJ</i> , 710, L92 | Acciari, V. A. et al. (VERITAS collaboration) 2010, <i>ApJ</i> , 474, 937 |
| Abdo, A. A. et al. (Fermi collaboration) 2011, <i>ApJ</i> , 734, 28 | Acero, F. et al. 2009, <i>Astron Astrophys</i> , 505, 157 |
| Abramowsky, A. et al. (HESS collaboration) 2011, <i>Astron Astrophys</i> , 531, 81 | Accero, F. et al. (HESS collaboration) 2010, <i>Astron Astrophys</i> , 516, 62 |

- Adriani, O. et al. (Pamela collaboration) 2011a, *Science*, 458, 607
- Adriani, O. et al. (Pamela collaboration) 2011b, *Science*, 332, 69
- Albert, J. et al. (Magic collaboration) 2007, *ApJ*, 664, L87
- Aharonian, F. A. 2004, *Very High Energy Cosmic Gamma Radiation: A Crucial Window on the Extreme Universe (River Edge: World Scientific)*
- Aharonian, F. A., & Atoyan, A. M. 1999, *Astron Astrophys*, 351, 1999
- Aharonian, F. A., & Sunyaev, R. A. 1984, *MNRAS*, 210, 257
- Aharonian, F. A., Drury, L. O'C., Völk, H. J. 1994, *Astron Astrophys*, 285, 645
- Aharonian, F. A., et al. (HESS collaboration) 2004, *Nature*, 432, 75
- Aharonian, F. A. et al. (HESS collaboration) 2006, *Astron Astrophys*, 449, 223
- Aharonian, F. A. et al. (HESS collaboration) 2007, *Astron Astrophys*, 464, 235
- Aharonian, F. A., Buckley, J., Kifune, T., & Sinnis, G. 2008a, *Rep Prog Phys*, 71, 096901
- Aharonian, F. A. et al. (HESS collaboration) 2008b, *Astron Astrophys*, 481, 401
- Aharonian et al. (HESS collaboration) 2009, *ApJ*, 692, 1500
- Aharonian, F. A. et al. (HEGRA collaboration) 2010, *A&A*, 370, 112
- Albert, J. et al. (MAGIC collaboration) 2010, *ApJ*, 474, 937
- Atoyan, A. M., Tuffs, R., Aharonian, F. A., & Völk, H. J. 2000a, *A&A* 354, 915
- Atoyan, A. M., Aharonian, F. A., Tuffs, R., & Völk, H. J. 2000b, *A&A* 355, 211
- Anderson, M., Rudnick, L., Leppik, P., Perley, R., & Braun, R. 1991, *ApJ*, 373, 146
- Baade, W., & Zwicky, F. 1943, *Phys Rev*, 46, 67
- Bamba, A., Yamazaki, R., Keno, M., & Koyama, K. 2003, *ApJ*, 589, 827
- Bell, A. R. 2004, *MNRAS*, 353, 550
- Berezhko, E. G., & Ellison, D. C. 1999, *ApJ*, 526, 385
- Berezhko, E. G., & Völk, H. J. 2010, *Astron Astrophys*, 511, id.A34
- Berezinsky, V. S., Bulanov, S. V., Ginzburg, V. L., Dogiel, V. A., Ptuskin, V. S. 1990, *Astrophysics of Cosmic Rays (Amsterdam; North-Holland)*
- Borkowski et al. 2010, *ApJ*, 724, 161
- Casanova, S. et al. 2010, *PASJ*, 62, 1127
- Cassam-Chenai, G., Decourchelle, A., Ballet, J. et al. 2004, *Astron Astrophys*, 427, 199
- Churazov, E., Sazonov, S., Tsygankov, S., Sunyaev, R., & Varshalovich, D. 2011, *MNRAS*, 411, 1727
- Chevalier, R. 1977, *Annu Rev Astron Astrophys*, 15, 175
- Cowsik, R., & Sarkar, S. 1984 *MNRAS*, 207, 745
- Drury, L. O'C., Aharonian, F. A., & Völk, H. J. 1994, *Astron Astrophys*, 287, 959
- Drury, L. O'C., Aharonian, F. A., Malyshev, D., & Gabici, S. 2009, *Astron Astrophys*, 496, 1
- Elisson, D., Patnaude, D. J., Slane, & Raymond, J. 2010, *ApJ*, 712, 287
- Fukui, Y. 2008, *AIP Conf Proc*, 1085, 104
- Gabici, S., & Aharonian, F. 2007, *ApJ*, 665, L131
- Gabici, S. et al. 2010, arXiv:1009.5291v1
- Gaisser, T. K. 1990, *Cosmic Rays and Particle Physics (Cambridge University Press)*
- Ginzburg, V. L., & Syrovatskii, S. I. 1964, *Origin of Cosmic Rays (New York: Macmillan)*
- Giuliani, A. et al. (Agile collaboration) 2010, *Astron Astrophys*, 516, L11
- Helder, E. A., & Vink, J. 2008, *ApJ*, 686, 1094
- Helder, E. A. et al. 2009, *Science*, 325, 719
- Hinton, J. A., & Hofmann, W. 2009, *Annu Rev Astron Astrophys*, 47, 523
- Iyudin, A. F. et al. 1994, *Astron Astrophys*, 284, L1
- Katz, B., & Waxman, E. 2008, *JCAP*, 1, 18
- Kelner, S. R., Aharonian, F. A., & Bugayov, V. V. 2006, *Phys Rev D*, 74, 034018
- Kirk, J. G., Duffy, P., & Gallant, Y. A. 1996, *Astron Astrophys*, 314, 1010
- Koyama, K. et al. 1995, *Nature*, 378, 255
- Kozlovsky, B., Murthy, R., & Ramaty, R. 2002, *ApJS*, 141, 523
- Lagage, P. O., & Cesarsky, C. J. 1983, *Astron Astrophys*, 125, 249
- Lazendic, J. S., Slane, P. O., Gaensler, B. M. et al. 2004, *ApJ*, 602, L201
- Malkov, M. A., & Drury, L. O'C. 2001, *Rep Prog Phys*, 64, 429
- Malkov, M. A., Diamond, P. H., & Sagdeev, R. Z. 2005, *ApJ*, 624, L37
- Michelson, P. F., Atwood, W. B., & Ritz, S. 2010, *Rep Prog Phys*, 73, 074901
- Morlino, G., Amato, E., & Blasi, P. 2008, *MNRAS*, 392, 240
- Morrison, P. 1957, *Nuovo Cimento*, 7, 858
- Montmerle, T. 1979, *ApJ*, 231, 95
- Paul, J., Casse, M., & Cesarsky, C. J. 1976, *ApJ*, 207, 62
- Prantzos, E. et al. 2011, *Rev Mod Phys* 83, 1001
- Renaud, M. et al. 2006, *ApJ*, 647, L41
- Slane, P. et al. 1999, *ApJ*, 525, 357
- Stephenson, F. R., & Green, D. A. 2001, *Historical Supernovae and Their Remnants (Oxford: Clarendon Press)*
- Tanaka, T., Uchiyama, Y., Aharonian, F. et al. 2008, *ApJ*, 685, 988

- Tavani, M. 2010, *Nucl Instrum Methods Phys Res A*, 630, 7
- Tavani, M. et al. (Agile collaboration) 2010, *ApJ*, 710, L151
- Torres, D. F., Marrero, A. Y. R., & de Cea Del Pozo, E. 2010, *MNRAS*, 408, 1257
- Uchiyama, Y., & Aharonian, F. A. 2008, *ApJ*, 677, L195
- Uchiyama, Y., Aharonian, F., Takahashi, T. et al. 2003, *Astron Astrophys*, 400, 567
- Uchiyama, Y., Aharonian, F., Tanaka, T., Takahashi, T., & Maeda, Y. 2007, *Nature*, 449, 576
- Villante, F. L., & Vissani, F. 2007, *Phys Rev D*, 76, 125019
- Vink, J., Yamazaki, R., Helder, E. A., & Schure, K. M. 2010, *ApJ*, 722, 1727
- Völk, H. J., Berezhko, E. G., & Ksenofontov, L. T. 2003, *Astron Astrophys*, 409, 563
- Wang, Z. R., Qu, Q., & Chen, Y. 1997, *Astron Astrophys*, 318, L59
- Zirakashvili, V. N., & Aharonian, F. A. 2007, *Astron Astrophys*, 465, 695
- Zirakashvili, V. N., & Aharonian, F. A. 2010, *ApJ*, 708, 965
- Zirakashvili, V. N., & Aharonian, F. A. 2011, *Phys Rev D*, 84, 8
- Zirakashvili, V. N., Aharonian, F. A., Ona di Wilhelmi, E., & Tuffs, R. 2011, submitted to *ApJ*

16 Galactic Distance Scales

Michael W. Feast

Astronomy Department and Astrophysics, Cosmology and Gravity
Centre, University of Cape Town and South African Astronomical
Observatory, Rondebosch, South Africa

1	<i>Introduction and Basic Methods</i>	831
1.1	Trigonometrical Parallaxes	831
1.2	Statistical Parallaxes	832
1.3	Pulsation Parallaxes	833
1.4	Using Galactic Rotation for Distance Calibration	834
2	<i>Important Distance Indicators</i>	835
2.1	Classical Cepheids	835
2.2	Type II Cepheids	839
2.3	RR Lyraes Variables	841
2.4	Miras Variables	844
2.4.1	Zero Point for O-Miras	845
2.4.2	Zero Point for C-Miras	846
2.5	δ Sct and SX Phe Variables	847
2.6	Red Clump	847
2.7	Tip of Red Giant Branch	848
2.8	Novae	850
2.9	Eclipsers	850
2.10	Spectroscopic Parallaxes	850
2.10.1	The MK System	851
2.10.2	OB Stars and Supergiants	851
2.10.3	The Most Luminous Supergiants	852
2.11	Large-Scale Studies of Common Stars	852
2.12	Open Clusters	853
2.13	Globular Clusters	854
3	<i>The Scale of Our Galaxy</i>	854
3.1	Distance to the Galactic Center	854
3.1.1	Orbits Around the Central Black Hole	854
3.1.2	Miras	855
3.1.3	Red Clump	856
3.1.4	Type II Cepheids	856
3.1.5	RR Lyraes	856
3.1.6	Kinematic Determination	856
3.1.7	Parallax of Sgr B2	857
3.1.8	Classical Cepheids	857

3.1.9	Summary	858
3.2	The Scale Length of the Galactic Disc	858
4	<i>Comparison of Distance Scales</i>	859
4.1	NGC4258	859
4.2	LMC	859
4.2.1	Classical Cepheids	859
4.2.2	Type II Cepheids	860
4.2.3	RR Lyraes	860
4.2.4	Eclipsers	861
4.2.5	Miras	861
4.2.6	SN1987A	861
4.2.7	Red Clump	862
4.2.8	Tip of the Red Giant Branch	862
4.2.9	δ Sct Stars	862
4.2.10	Summary	862
4.3	The Fornax Dwarf Spheroidal	862
4.3.1	TRGB(K)	863
4.3.2	TRGB(I)	863
4.3.3	RR Lyrae Variables	863
4.3.4	Red Clump	864
4.3.5	Miras	864
4.3.6	δ Sct Stars	864
4.3.7	Summary	864
5	<i>Conclusions</i>	864
	<i>Appendix: Bias Correction</i>	864
A.1	Introduction	865
A.2	Correction of Bias in Distance Moduli	865
A.3	Bias in Distances from Absolute Magnitudes	867
A.4	Bias in Mean Parallaxes and Absolute Magnitudes	868
A.5	Determination of Absolute Magnitudes: The Reduced Parallax Method	869
A.6	General Comments	871
	<i>Acknowledgements</i>	871
	<i>References</i>	871

Abstract: This chapter begins with a discussion of the basic methods of determining astronomical distances, particularly, trigonometrical, statistical, and pulsational parallaxes. It then summarizes the current state of the calibration of various classes of pulsating variables (Classical Cepheids, type-II Cepheids, RR Lyraes, Miras, and δ Sct and SX Phe stars). Work on other distance indicators (e.g., the red giant clump and the tip of the red giant branch) is also summarized. The use of spectroscopic parallaxes and their application to supergiants and common stars as well as the methods of determining the distances to open and globular clusters are discussed. To illustrate and compare different distance indicators, their use in estimating the scale length of our Galaxy, and the distance to Galactic centre as well as the distances to the LMC, the Fornax dwarf spheroidal, and the spiral galaxy NGC4258 is discussed in some detail. An appendix summarizes some common bias problems that arise in the calibration and use of distance indicators.

1 Introduction and Basic Methods

An understanding of the structure of our own and other galaxies rests ultimately on the establishment of sound distance scales. Whether one is interested in the space densities, luminosities, and motions of stars in the solar neighborhood, the large-scale structure of the Galaxy or the composition, structure, and dynamics of other galaxies, distances need to be found. In some cases, distances of objects can be estimated directly. However, more generally, it is necessary to estimate luminosities for various types of object from which distances to objects of the chosen class can be derived.

In this chapter, various distance indicators currently in use, are discussed in some detail. At the present time, the calibration of distance indicators is a very active and evolving field of research. However, the types of indicator and the problems encountered in calibrating them will remain, though numerical values change.

For studies of the large-scale structure of our Galaxy, two distances are of special importance, the distance to the Galactic Centre, R_o , and the scale length of the Galactic disc, R_d , and estimates of these quantities are discussed. Finally, to further compare different distance indicators and to illustrate in detail some of the problems which arise in the use of distance indicators, a few galaxies whose distances have been estimated in different ways are discussed.

In establishing and using distance scales, the problem of bias in the samples used has to be considered. In some cases, bias, if not corrected for, can lead to substantial systematic errors. An appendix summarizes likely bias problems in distance estimation, and this is referred to at various places in the text.

1.1 Trigonometrical Parallaxes

In principle, trigonometrical parallaxes provide the soundest, assumption free, estimates of stellar distances. However, parallaxes, at least of interesting distance indicators, tend to be very small. Ground-based parallax measurements have in the past been plagued by a variety of systematic problems (see, e.g., Strand 1963), though some of these have been overcome (see summary in Perryman 2009, especially Section 2.11).

A major step forward has been the ability to measure parallaxes of very high precision with the fine guidance sensors on HST. However, the HST work, like the ground-based work, leads to parallaxes relative to a set of stars close to each programme star. Thus, the measured parallaxes have to be converted to absolute using estimates of the (true) parallaxes of the comparison stars. This is done either by assuming a distribution of stars based on some Galactic model or using spectroscopy and photometry to estimate the parallaxes of the comparison stars. The procedure is discussed in detail by Benedict and McArthur (2004). Especially when the programme star's parallax is small, the main uncertainty in the final result comes from the correction from relative to absolute parallaxes. For instance, in the case of ten Cepheid parallaxes measured using the HST (Benedict et al. 2007), the average estimated true parallax of the comparison stars is about one third of the measured relative parallax of the Cepheid. The uncertainty introduced by this is, of course, taken into account in estimating the uncertainty of the final absolute Cepheid parallaxes.

The Hipparcos satellite transformed all fields of astrometry (see Perryman 2009). The positions, proper motions, and parallaxes of about 118,000 stars were determined from a global solution based on the positions of radio galaxies, and the system of parallaxes can be taken as absolute with a very small uncertainty. More than 20,000 of the stars had their distances measured with an uncertainty of less than 10%. A complete rediscussion of the original data has been carried out by van Leeuwen (2007), resulting in a considerable improvement of the parallaxes, at least for the brighter stars.

There are a number complications which may arise in the case of both relative and absolute parallaxes. For instance, if the star is an unresolved binary, the observed primary may well move in its orbit during the period covered by the observations. In general, a good series of observations will show the effect of this when a solution is made for parallax and proper motion (the latter must, of course, always be derived together with the parallax). If the angular diameter of a star is of comparable size to its parallax, as, for instance, in the case of Mira variables (see [Sect. 2.4](#)), spurious results might result from changing nonuniformity (e.g., spots) on the stellar surface.

Some future ground and space-based parallax projects are mentioned in [Sect. 2.11](#), and the measurement of parallaxes at radio frequencies using VLBI is discussed in [Sect. 2.4](#).

1.2 Statistical Parallaxes

If a homogeneous population of stars all of the same absolute magnitude has a bulk mean velocity relative to the Sun, an analysis of both the proper motions and radial velocities gives the direction of this motion. The radial velocities give the mean velocity directly. The proper motions give a distance-dependent mean velocity and the apparent magnitudes give relative distances. Thus, the absolute magnitude (a distance scale) of the stars can be estimated. Similarly, a comparison of the velocity dispersion of the stars derived from radial velocities and proper motions also leads to a distance scale (see, e.g., Mihalas 1968).

In the older literature, results were obtained separately from the mean motion (secular parallax) and from the velocity dispersion(s) (statistical parallax). It is now usually to solve for a distance scale using bulk motions and dispersions together. This joint procedure is now referred to as the method of statistical parallaxes. Maximum likelihood procedures have been developed following the general principles set out by Murray (1983). These essentially assume some general Galactic model with, for instance, the velocity dispersions depending on direction in the

Galaxy. Several realizations of the method are summarized by Layden (1999), and one method has been discussed in considerable detail by Popowski and Gould (1998a, b, c). The method is, in principle, particularly suitable for objects with large mean motion relative to the Sun and a high-velocity dispersion, such as objects in the Galactic halo. However, it does depend on the adopted Galactic model.

The moving cluster method to determine the distance of a group of stars is simply a form of the secular parallax solution applied to a group of stars all of which are believed to have a common space motion and are spread over a significant area of the sky, e.g., the Hyades cluster. Refinements to the method are summarized by Perryman (2009 esp. Chapter 6). An extension, which they term the method of Galactic parallax, has been proposed by Eyre and Binney (2009). They show how the proper motions of stars in a tidal stream moving on an orbit in the Galactic force-field together with the velocity of the Sun about the Galactic centre can be used to obtain distances to points in the stream.

1.3 Pulsation Parallaxes

Parallaxes of pulsating stars can be estimated using their magnitude, color, and radial velocity variations. The method of pulsation parallaxes is often known as the Baade–Wesselink method or the surface brightness method. In modern work, it has been usual to use the surface brightness approach (see summary in Feast and Walker 1987).

It is assumed that the surface brightness of the pulsating star is a function of the color and that this can be calibrated, either from nearby, constant stars of known distance or theoretically. Variations in color and magnitude during the pulsation cycle then give the variations in angular radius of the star. Radial velocity observations give the radius changes in linear measure and thus lead to a distance estimate.

The method in principle provides an independent way to the distance scale for pulsating stars. However, in practice, there are a number of complications. The relation between surface brightness and color is, in general, not unique, depending, for instance, on surface gravity which changes with changing radius. A detailed discussion of this issue is given by Laney and Stobie (1995). The effect is minimized by working with magnitudes and colors in the near infrared (generally either $K, J - K$ or $K, V - K$ have been used).

The uncertainties in the derivation of the surface brightness is being successfully bypassed in the case of Cepheids by direct interferometric measurements of the angular diameter of the star and its variation with phase (Lane et al. 2002 and references there).

As regards the pulsation velocities, there are at least two potential problems:

1. It is assumed that the absorption lines being measured for velocity are produced by material at the radii estimated from the photometry or measured interferometrically. However, it is known, for instance, that high-resolution work on Cepheids shows gradients of velocity with ionization/excitation potential (e.g., Wallerstein et al. 1992). Also some variables (e.g., RR Lyraes, some type II Cepheids) show doubling of absorption lines at some phases or discontinuities in the radial velocity curves due to shock waves in their atmospheres and evidence for two separate absorbing layers. Such phases must be omitted in solutions.
2. The measured radial velocity is an integrated value over the whole visible hemisphere of the star. To convert it to the velocity seen radially from the centre of the star, it has to be

multiplied by a projection factor (p). The value of p is obviously affected by the limb darkening. It may be possible to measure this directly in the future by interferometry. Uncertainty, in p is discussed further in the case of Cepheids in [◆ Sect. 2.1](#) and is probably the main current uncertainty in this method for these stars.

The determination of the distances of eclipsing binaries by combining light, color, and radial velocity curves has some similarity to the pulsation parallax method in that it requires the estimation of the surface brightness in order to obtain a distance. This method is discussed in [◆ Sect. 2.9](#) and in connection with the distance to the LMC ([◆ Sect. 4.2](#)).

Another similar method (Hendry et al. 1993) is to estimate a stellar radius by combining a spectroscopically measured rotational velocity (line widths) with a rotation period determined photometrically from periodicity due to star spots. This can then be combined with surface-brightness estimates. Since the measured quantity is the projected rotational velocity, $v \sin i$, the method is suitable for groups of stars (e.g., a cluster) in which the distribution of the inclinations of the rotation axes can be considered random and the effect of the $\sin i$ term can be statistically estimated.

1.4 Using Galactic Rotation for Distance Calibration

For objects which have, in the mean, circular orbits about the Galactic center, it is possible to use the effects of differential Galactic rotation directly to obtain a distance scale if the stars are spread out over a significant volume. Analysis of the kinematics of such objects yields Galactic constants particularly the Oort constants A and B , where,

$$A = -\frac{1}{2}R_o(d\omega/dR)_o \quad (16.1)$$

and

$$A - B = \omega_o = \Theta_o/R_o \quad (16.2)$$

Here, R is the distance of the object from the Galactic center, whilst Θ is the Galactic circular velocity at that point, and $\omega = \Theta/R$ is the angular velocity. The subscript, o , denotes the values at the position of the Sun.

The equations relating these quantities to measured proper motions and radial velocities are to a first order:

$$\kappa\mu_{l*} = (u_o \sin l - v_o \cos l)/r + (A \cos 2l + B) \cos b \quad (16.3)$$

and

$$V_r = -u_o \cos l \cos b - v_o \sin l \cos b - w_o \sin b + Ar \sin 2l \cos^2 b. \quad (16.4)$$

Here b, l are the Galactic latitude and longitude; u_o, v_o, w_o are the components of the solar motion relative to the local standard of rest (the circular velocity at the Sun) directed toward the Galactic center, in the direction of Galactic rotation, and perpendicular to the Galactic plane. $\mu_{l*} = \mu_l \cos b$ and μ_l is the proper motion in Galactic longitude, V_r is the radial velocity, and r is the distance of the object from the Sun. If a proper motion μ is in milliarcsecs and distances r in kiloparsecs, then the velocity $\kappa\mu r$ is in km s^{-1} , where $\kappa = 4.74$.

Higher-order terms can be added to these equations in both proper motions and radial velocities. The Oort constants derived from the proper motions are very insensitive to the distance scale assumed. Feast and Whitelock (1997) found $A = 14.82 \pm 0.84 \text{ km s}^{-1} \text{ kpc}^{-1}$,

$B = -12.37 \pm 0.64 \text{ km s}^{-1} \text{ kpc}^{-1}$ from Hipparcos proper motions of Cepheids.¹ On the other hand, as [16.4](#) shows, radial velocities yield the quantity Ar and with a known A can be used to derive a distance.

2 Important Distance Indicators

2.1 Classical Cepheids

Classical Cepheids have long been thought to hold the key to accurate determinations of distances within our own Galaxy and to other galaxies. Hence, a vast amount of effort both observational and theoretical has gone into attempts to refine our understanding of these variables and their absolute magnitudes following the discovery of a period-luminosity (PL) relation in the Small Magellanic Cloud from photographic data (Leavitt and Pickering 1912).

With periods in the range 1 to about 100 days, amplitudes of typically ~ 0.5 mag and distinctive light curves (see, for instance, Sterken and Jaschek 1996) they are easily recognized at large distances. Most pulsate in the fundamental mode, but there are first overtone pulsators and a small fraction of pulsators in higher or mixed modes. Magellanic Cloud observations show that the overtone pulsators have PL relations displaced from the fundamental PL. The overtones are generally recognizable from their lower amplitudes and near-sinusoidal light curves, but some confusion is possible. There is also a small chance of confusion with type-II Cepheids (see [Sect. 2.2](#)), though these generally have distinctive light curves.

The early history of the study and calibration of the PL relation is given by Fernie (1969) (see also van den Bergh 1975). A series of important investigations by Sandage and Tammann was summarized by Sandage (1972). The more recent development of the subject may be followed in various reviews (e.g., Feast and Walker 1987; Feast 1999b; Sandage and Tammann 2006, and in a number of conference volumes, for instance, Alloin and Gieren 2003). These will show that complete agreement has not always been obtained on some aspects of this topic.

An important step in the understanding of these stars was the recognition by Sandage (1958) that a PL relation (e.g., $M_V = A \log P + B$) has a finite width due to the width of the instability strip and that this might be reduced using a color term. That is, using a period-luminosity-color (PLC) relation (e.g., Sandage and Tammann 1969). Thus,

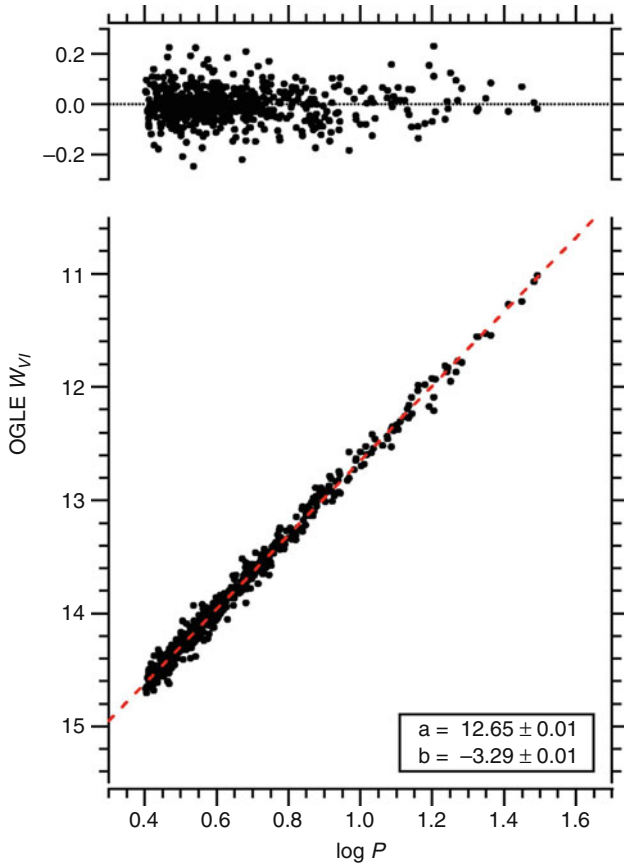
$$M_V = \alpha \log P + \beta_1(V - I)_o + \gamma \quad (16.5)$$

It is then easily seen that the distance modulus ($\text{Mod} = V_o - M_V$) is given by,

$$\text{Mod} = W_{VI} - \alpha \log P - \gamma - (\beta_1 - \beta_2)(V - I)_o \quad (16.6)$$

Where $W_{VI} = V - \beta_2(V - I)$ and $\beta_2 = A_V/E(V - I)$ is the ratio of total to selective absorption. Reddening corrected values are denoted by subscript o whilst measured values are unsubscripted. Work on LMC Cepheids by the OGLE group (Udalski et al. 1999) showed that $\beta_1 \sim \beta_2$ so that the last term in [16.6](#) is negligible and any scatter in the relation between W_{VI} and $\log P$ in, for instance, the LMC, should be small. That this is the case is shown in [Fig. 16-1](#)

¹Miyamoto and Zhu (1998) obtained somewhat larger values of A from Hipparcos proper motions of both OB stars and Cepheids. This may be at least partly due to the fact that the velocity dispersions of the stars do not seem to have been taken into account in their weighting procedure.



■ Fig. 16-1

The relation $W_{VI} = a + b(\log P - 1.00)$ for 581 LMC OGLE Cepheids (Benedict et al. 2007, reproduced by permission of the AAS). Residuals are shown at the top

and in the following the final term in \blacktriangleright 16.6 is omitted. If $(\beta_1 - \beta_2)$ is significant, a scatter is introduced into the $W_{VI} - \log P$ relation by the spread in $(V - I)_o$ at a given period.

The LMC PL relation in V shows an apparent change in slope at $P \sim 10$ days (e.g., Ngeow et al. 2005). Whether this is real or an effect of systematic errors in reddening corrections is not entirely clear. However, there is good evidence (Ngeow and Kanbur 2005) that the relation between W_{VI} and $\log P$ is closely linear.²

There has been much discussion as to the dependence of the slope α and the zero-point γ of \blacktriangleright Eqs. 16.5 and \blacktriangleright 16.6 on metallicity.

The situation as regards α is set out in \blacktriangleright Table 16-1. Various values of β_2 have been used in forming this table corresponding to different assumed reddening laws. However, the range in

²The procedure followed by the Sandage group (e.g., Saha et al. 2006) is to use observed PL relations at V and I separately and to combine them to obtain reddening-free distance moduli. This is obviously equivalent in principle to using a relation in W_{VI} .

■ **Table 16-1**

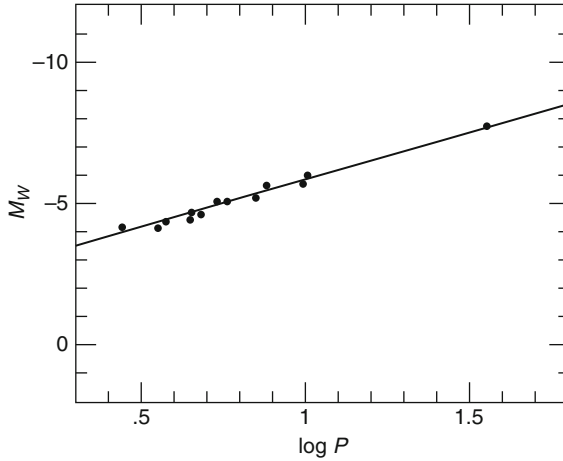
Slope, α of the PLC(V,I) relation

No	Object	α	β_2	Method	Ref.
1	LMC	-3.277 ± 0.014	2.55		Udalski et al. 1999
2	LMC	-3.29 ± 0.01	2.45		Benedict et al. 2007
3	SMC	-3.303 ± 0.022	2.55		Udalski et al. 1999
4	SMC	-3.277 ± 0.023	2.45		Di Benedetto 2008
5	NGC4258	-3.18 ± 0.13	2.45		van Leeuwen et al. 2007
6	Galaxy	-3.29 ± 0.15	2.45	Trig. Par.	van Leeuwen et al. 2007
7	Galaxy	-3.46 ± 0.07	2.55	Puls. Par.	Fouqué et al. 2007
8	Galaxy	$-3.74 \pm ?$	2.52	Clusters etc	see text

these is not sufficient to affect significantly a comparison of the values of α . The value for a region of near solar metallicity in the galaxy NGC4258 is derived from the data of Macri et al. 2006. The directly determined slopes in the three galaxies (SMC, LMC, and NGC 4258) of increasing metallicity all agree within the uncertainties. The three entries for our own Galaxy illustrate the situation there. Item 6 in [Table 16-1](#) is from HST and Hipparcos trigonometrical parallaxes, and this is illustrated in [Fig. 16-2](#). There is an interesting problem in this case concerning the question of bias corrections for the parallaxes. The corrections actually applied were for a standard “Lutz-Kelker”-type bias (see [Sect. A.4](#)). This applies for objects selected by parallax only. In fact, the Cepheids selected for parallax measurement were chosen at least partly on the basis of being those of the brightest apparent magnitude for their period. In principle, this introduces a “Malmquist”-type bias (see [Sect. A.2](#)) which (in this case) would be of opposite sign to the Lutz–Kelker bias. In the present case, any correction is small and similar from star to star so that the result for the slope is only marginally affected. However, the complexities resulting from selection criteria (often poorly known) are likely to occur in many distance scale problems. As is clear from [Fig. 16-2](#), the estimate of the uncertainty in the slope depends crucially on the one long period Cepheid (*l* Car), and this remains a weakness in the result until more high-quality trigonometrical parallaxes of long-period Cepheids are obtained.

The value of α from Fouqué et al. (2007) (item 7) includes trigonometrical parallaxes but depends mainly on pulsation parallaxes. As they point out, the result depends rather critically on the relation adopted for the dependence of the projection factor p (see, [Sect. 1.3](#)) on period, and there is no strong evidence from their result for a difference in slope between the LMC and our Galaxy.

The final entry in the table, which suggests a steeper value of α , comes from the work of Sandage and Tammann summarized in Saha et al. (2006). It is based on some earlier pulsation parallaxes, which now fall away in view of later work, and data from Cepheids in open clusters. There are a number of clusters containing Cepheids with periods shorter than ~ 10 days whose distances seem well established (e.g., An et al. 2007), and these give a value of α similar that found in the LMC. A steeper value is indicated by clusters containing Cepheids of longer period (see Fouqué et al. 2007, Figure 1). However, there is a distinct difference in the nature of the data at the longer and shorter periods. At the shorter periods, the cluster distances are mainly from main-sequence fitting of well-defined clusters (see [Sect. 2.12](#)). At longer periods, the results are primarily from stellar associations or poorly defined clusters with distances coming mainly from spectral classification of presumed members. In a number of cases, the membership of



■ Fig. 16-2

$M_W = M_V - 2.45(V - I)$ for Cepheids with good trigonometrical parallaxes plotted against $\log P$. The line is from [Eq. 16.7](#) (Reproduced from van Leeuwen et al. (2007) MN 379, 723)

the Cepheid in the cluster is uncertain and, in general, it is not clear how closely the spectral type – luminosity scale used is tied to the the main sequence fitting scale. Thus, these results cannot at this stage be considered definitive. The situation is unsatisfactory in that it is often the longer-period stars which are of importance for distance-scale problems.

In the following, it is assumed that the $W_{VI} - \log P$ slope in our Galaxy is the same as that in the LMC but, clearly, further work is needed on this matter.

A combination of HST and Hipparcos parallaxes by the reduced parallax method (see [Sect. A.5](#)) leads to

$$M_W = -3.29 \log P - 2.58(\pm 0.03) \quad (16.7)$$

where M_W is the absolute magnitude corresponding to the quantity W_{VI} .

Similar results to the above apply at near-infrared wavelengths. Here the effect of the width of the instability strip is less and can be ignored for many purposes. van Leeuwen et al. (2007) find,

$$M_K = -3.258 \log P - 2.40(\pm 0.05). \quad (16.8)$$

The slope is from the LMC (Persson et al. 2004) converted to the SAAO system.³

The dependence of the zero point (γ of [Eqs. 16.5](#) and [16.6](#)) on metallicity remains uncertain. Estimates have been made from galaxies where Cepheids have been observed in regions of different mean metallicity as determined from analysis of spectra of HII regions. In NGC 4258, Macri et al. (2006) found the effect to be $-0.49 \pm 0.15 \text{ mag dex}^{-1}$.⁴ On the other hand, the recent work by Benedict et al. (2011) (see [Sects. 2.3](#) and [4.2](#)) indicates that a Cepheid distance to the LMC based on the above Galactic calibration, without any metallicity

³In the case of near-infrared photometry, there are appreciable differences between different filter systems (see, e.g., Carpenter 2001).

⁴This is in the “ T_e ” abundance system. In application, it is essential to ensure that adopted metallicities are as close to this system as possible.

correction agrees well with an RR Lyrae based distance. For the purposes of the present chapter, it will be assumed that metallicity effects on Cepheids (at least between Galactic calibrators and the LMC) are negligible. However, the possibilities of (small) metallicity effects should be borne in mind. Such effects might well be nonlinear.

In the Magellanic Clouds and elsewhere, there are a small number of Cepheids with periods near 100 days. These have generally been omitted from distance-scale studies because they do not fit PL relations as well as those of shorter period and have less regular periods. They may be useful as distance indicators since they are the most luminous Cepheids (e.g., Bird et al. 2009).

2.2 Type II Cepheids

Type II Cepheids (CephIIs) have periods in the same range as classical Cepheids but are older, low-mass stars which are found in globular clusters of a range of metallicities and also in the halo and old disc populations. Their light curves are distinctive (Sterken and Jaschek 1996). They are believed to be passing through an instability strip either on their way from the horizontal branch to the AGB, on blue loops from the AGB, or during their final exit from the AGB (see Gingold 1985). Conventionally, they are divided by period into three groups, which may be related to the three modes of evolution just mentioned. These groups are: BL Her(BL) stars ($P \leq 4$ days), W Vir (WV) stars (P from 4 to 20 days), and RV Tau (RV) stars ($P > 20$ days). These period limits are not too well defined.

Early work on these stars suggested period-luminosity relations at optical wavelengths (see summary in Pritzl et al. 2003), though the nature of these relations and their linearity and possible dependence on metallicity remained uncertain. A number of recent studies have clarified the position considerably, and it is now possible to see rather clearly the potential, and limitations, of these stars as distance indicators.

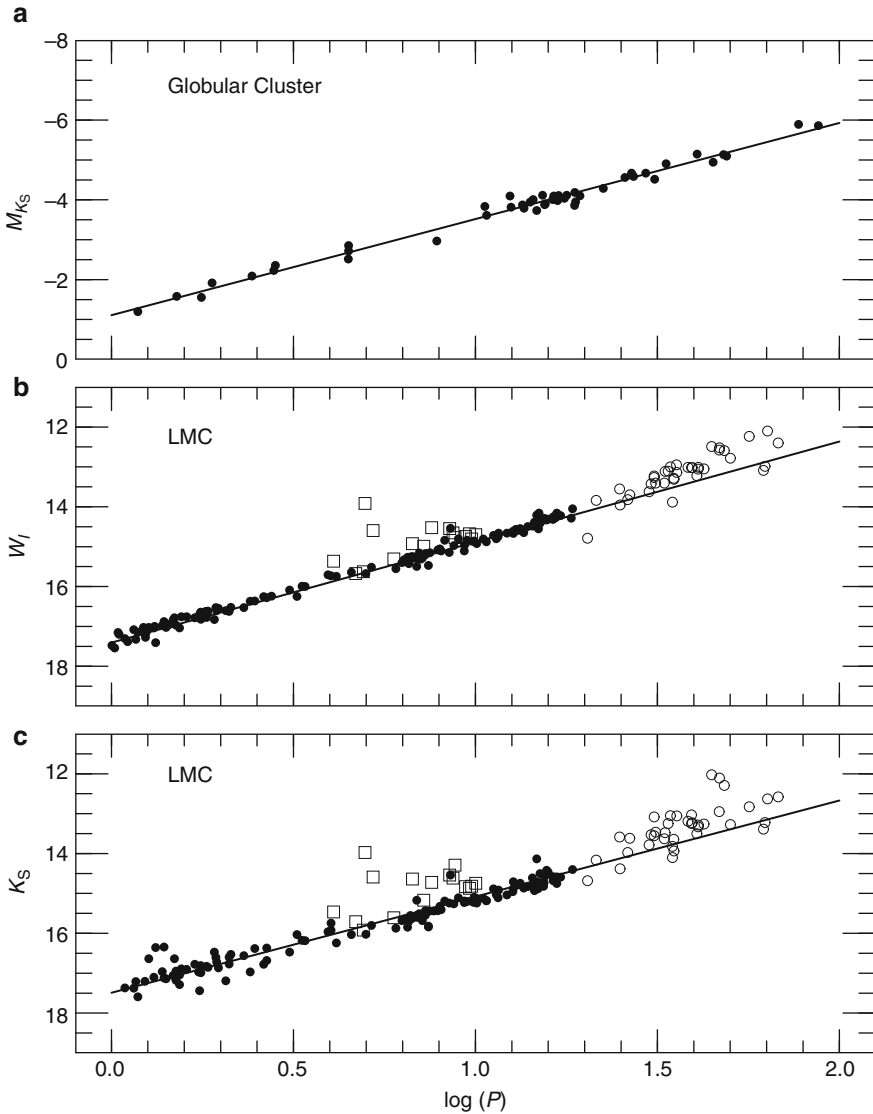
Matsunaga et al. (2006) obtained JHK_s light curves of 46 CephIIs in 26 globular clusters.⁵ The relative distances of the clusters was determined from their RR Lyrae or HB stars and assuming

$$M_V(HB) = 0.22[\text{Fe}/\text{H}] + \gamma \quad (16.9)$$

see [Sect. 2.3](#). [Figure 16-3a](#) shows the derived results at K_s . These indicate a linear relation over the whole period range of CephIIs. The scatter about the relation is small, 0.14 mag, and this together with the fact that the CephIIs in some clusters cover a range of periods, strongly suggests that any metallicity correction to the mean relation will be small.

The OGLE III survey of the LMC yields important results on CephIIs. (Soszyński et al. 2008). There is considerable scatter and nonlinearity for PL relations in V and I . The quantity $W_I = I - 1.55(V - I)$ shows lower scatter, and the shorter-period stars (BL and WV) show a narrow linear relation (see [Fig. 16-3b](#)). The longer period (RV) stars show a wider scatter and deviate in the mean from this linear relation. It should be noted that while the quantity W_I is defined to be reddening free, it also corrects, at least to some extent, for a real spread of color at a given period, that is, the existence of a true PLC relation. This correction would be nearly complete if the coefficient of the color term was similar to that for classical Cepheids (see [Sect. 2.1](#)). Unless the LMC CephIIs are a much more homogeneous group as regards metallicity than the Galactic ones, the low scatter in PL relations in W_I and K at the shorter periods (see below) indicates that the effect of metallicity on these relations is small.

⁵The photometry is in the system of the IRSF survey (Kato et al. 2007) which is close to the 2MASS system.



■ Fig. 16-3

PL relations for type II Cepheids. In b and c, filled and open circles are for objects with periods below and above 20 days. The open squares are for peculiar W Vir stars. See text for discussion

There are a number of stars classed as BL stars which lie near the anomalous Cepheid stars.⁶ In the WV period range, there are stars lying above the main body of WV variables and between them and the classical Cepheids. These stars are called peculiar WV (pW) stars

⁶Anomalous Cepheids which lie between the Classical and type II Cepheids in PL diagrams have not, so far, been studied in detail as possible distance indicators.

by Soszyński et al. They can be distinguished by the shape of their light curves and most, if not all, must be binaries since a significant number show eclipses.

• *Figure 16-3c* shows a K_s period-luminosity relation for the same OGLE LMC stars (Matsunaga et al. 2009a). There is no significant difference in PL slopes for the shorter period stars (<20 days) from that found for globular clusters. However, as • *Fig. 16-3c* shows, the longer-period stars, as a group, deviate from this relation. This is in marked contrast to stars of similar period in globular cluster (• *Fig. 16-3a*). Evidently, CepHII with periods greater than about 20 days are not suitable for use as distance indicators until this difference between clusters and the general LMC field is understood.

Unless one assumes a distance for the LMC or a globular cluster distance scale, the absolute calibration of the various Ceph II relations currently depends on the absolute magnitude of the two Galactic stars, V553 Cen ($\log P = 0.314$) and SW Tau ($\log P = 0.200$) derived from pulsation parallaxes (Feast et al. 2008). These lead to

$$M_{W_I} = -2.521(\log P - 1.2) - 4.12 \quad (16.10)$$

and

$$M_{K_s} = -2.410(\log P - 1.2) - 3.90 \quad (16.11)$$

where the slope of the W_I relation is taken from the LMC and that for the M_K relation from globular clusters (see Matsunaga et al. 2009a).

The uncertainty in these zero points is ~ 0.10 mag. The scatter about the relations is small, 0.10 for W_I in the LMC and 0.14 for M_K , part of which is observational, in globular clusters.

These results for the CephII stars may be summarized as follows. In globular clusters, there is a narrow linear PL relation in the near infrared extending over the whole period range of these stars ($\log P = 0.0$ to 2.0). In the general field, there may be other stars, perhaps younger, which lie above this PL (pW and RV Tau stars). Both these groups can probably be distinguished by their light curves.

2.3 RR Lyraes Variables

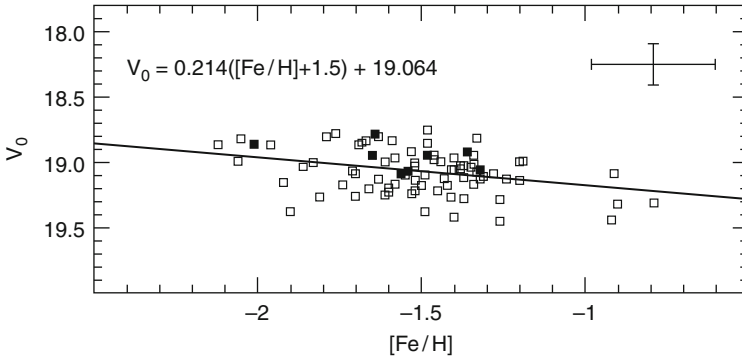
RR Lyrae variables are old, low-mass stars, pulsating with periods less than one day. They have been found in metal-poor globular clusters (at the intersection of the HB with an instability strip) in the halo and (old) disc of our Galaxy and in nearby galaxies. RRab stars are pulsating in the fundamental mode, RRC stars are first overtone pulsators and are of lower pulsation amplitude. With proper luminosity calibration, the RR Lyraes are of special importance for the distance scale of old populations, including globular clusters, dwarf spheroidal, and other nearby galaxies.

In a given globular cluster, the RR Lyraes all have nearly the same visual (V) magnitude independent of period, though the scatter increases with increasing metallicity (Sandage 1990). There is a long history of attempts to determine how M_V depends on metallicity, i.e., the value of α in the relation;

$$M_V = \alpha[\text{Fe}/\text{H}] + \beta, \quad (16.12)$$

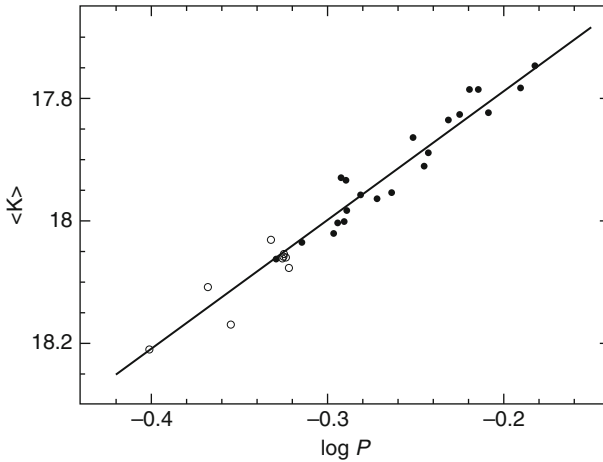
or, indeed, whether the relation is linear (for a summary, see Feast 1999a; McNamara 1999).

Possibly the best empirical determination comes from observations of RR Lyrae variables in the LMC (the general field not clusters) (Gratton et al. 2004) (see • *Fig. 16-4*). These give, $\alpha = 0.214 \pm 0.05$. This agrees with the slope found for Galactic RR Lyraes whose absolute



■ Fig. 16-4

The $V_0 - [\text{Fe}/\text{H}]$ relation for RR Lyrae variables in the LMC; open symbols for ab and c types, filled symbols for double mode pulsators (Gratton et al. 2004 A&A 421, 937, reproduced with permission ©ESO)



■ Fig. 16-5

The $K - \log P$ relation for RR Lyrae variables in the LMC cluster Reticulum. Open circles are for RRc variables after fundamentalization (Dall’Ora et al. 2004, reproduced by permission of the AAS)

magnitudes were determined from pulsation parallaxes, $\alpha = 0.20 \pm 0.04$, (see Fernley et al. 1998)⁷ and from globular clusters of different metallicity in M31, 0.13 ± 0.07 , (Fusi Pecci et al. 1996). However, Clementini et al. (2005) found a lower value in the Sculptor dwarf spheroidal galaxy ($\alpha = 0.09 \pm 0.03$).

An infrared $K - \log P$ relation was found in globular clusters by Longmore et al. (1986). ● Figure 16-5 shows an example for the cluster Reticulum in the LMC (Dall’Ora et al. 2004). A similar relation is found in the general field of the LMC (Szcwzyk et al. 2008) but with

⁷The value of α obtained in a consistent manner from pulsation parallaxes may well be satisfactory, though the derived absolute magnitudes may be affected by systematic errors (see below).

more scatter. Some of this is probably due to the finite depth of the LMC as well as a possible spread in metallicities. The full relation can be written,

$$M_K = \gamma \log P + \delta[\text{Fe}/\text{H}] + \phi. \quad (16.13)$$

Sollima et al. (2006) summarize the work on globular clusters and find $\gamma = -2.38 \pm 0.04$. The metallicity term is still uncertain. Theoretical work summarized by Sollima et al. suggests $\delta \sim 0.2$. An observational determination based on the relative distances of globular clusters of different metallicities gives $\delta = 0.08 \pm 0.11$ (Sollima et al.). Borissova et al. (2009) found $\delta = 0.05 \pm 0.07$ in the LMC. Thus, the metallicity effect may be quite small. Also, in our Galaxy the mean period decreases with increasing metallicity (see, e.g., Smith 1995, Fig. 1.5). Thus, to a first approximation, any metallicity dependence may be incorporated in the $\log P$ term.

Recently (Benedict et al. 2011), values of β and ϕ in \blacklozenge Eqs. 16.12 and \blacklozenge 16.13 have been obtained from HST trigonometrical parallaxes of five RR Lyrae variables. Using the method of reduced parallaxes (see \blacklozenge Sect. A.5).⁸ These results lead to

$$M_K = -2.38(\log P + 0.28) - 0.54(\pm 0.03) \quad (16.14)$$

and

$$M_K = -2.11(\log P + 0.28) + 0.05([\text{Fe}/\text{H}] + 1.58) - 0.54(\pm 0.03). \quad (16.15)$$

The zero points refer to the mean $\log P$ and metallicity of the HST sample. \blacklozenge Equation 16.14 adopts the $\log P$ slope as determined by Sollima et al. (2006) from globular clusters, and \blacklozenge Eq. 16.15 takes the $\log P$ and metallicity slopes from the work of Borissova et al. (2009) in the LMC. Also

$$M_V = 0.214([\text{Fe}/\text{H}] + 1.58) + 0.44(\pm 0.03). \quad (16.16)$$

These luminosities are considerably brighter than those obtained from statistical parallaxes (by ~ 0.3 mag in M_V (Popowski and Gould 1998a, b, c) and by ~ 0.4 mag in M_K (Dambis 2009)). The reasons for this are not entirely clear. It may indicate that the model of the Galactic halo adopted by these authors is too simplistic.

Pulsation parallaxes have been derived for number of RR Lyrae variables, but the results obtained depend on the stellar model adopted (see, for instance, Cacciari and Clementini (2003)). For the present, it appears best to base the RR Lyrae calibration on the trigonometrical parallaxes alone.

An important caveat, at least when using RR Lyrae variables to estimate distances of globular clusters, is the fact that in some relatively metal-rich globular clusters, there are RR Lyrae variables which are overluminous in M_V compared with the calibration discussed above (see, e.g., Matsunaga et al. 2009a; Pritzl et al. 2003). This anomaly, which is connected with the “second parameter” effect in globular clusters, has been much discussed and may be due to an overabundance of helium in these stars.

From the results in Carretta et al. (2000), the mean absolute magnitude of the horizontal branch (HB) has been taken as 0.03 mag brighter than \blacklozenge Eq. 16.16.

⁸Benedict et al. also give an alternative solution involving the application of a Lutz-Kelker bias correction (see \blacklozenge Sect. A.4).

2.4 Miras Variables

Mira variables are large amplitude ($\Delta V > 2.5$ mag.) stars at the end of their AGB evolution and may be either oxygen-rich (O-Miras) or carbon-rich (C-Miras). In the LMC, they show narrow period-luminosity relations at K from $\log P \sim 2.1$ to ~ 2.6 of the form,

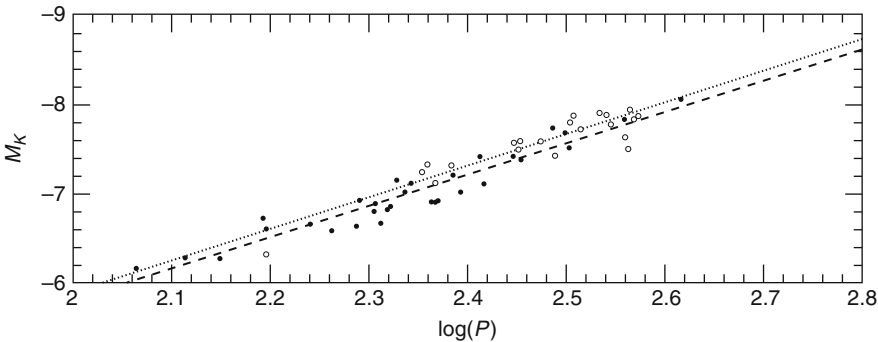
$$M_K = \alpha(\log P - 2.38) + \beta \quad (16.17)$$

Whitelock et al. (2008) (see \blacklozenge Fig. 16-6) and similar relations may be obtained in bolometric luminosities ($PL(M_{bol})$).⁹

At longer periods, the variables tend to be enshrouded in circumstellar dust, and some of them are OH/IR stars. Variables with thick dust shells tend to fall below the $PL(K)$ relation but remain close to a $PL(M_{bol})$ relation. Current evidence summarized in (Feast 2009) indicates that in the LMC the Miras are a mixture of O- and C-rich stars in range $\log P \sim 2.1$ to ~ 2.6 . From $\log P \sim 2.6$ to 3.0, they are mainly, if not all, C-stars, and above this period, they are O-rich. These compositional changes can be explained at least semi-quantitatively by nuclear processes and dredge-up. O-Miras with $\log P \sim 2.3$ are found in (metal-rich) globular clusters and evidently have an initial mass $< 1 M_\odot$. Evidence indicates that the initial mass rises as the period increases, being $\sim 4 M_\odot$ at $\log P \sim 3.0$. A similar age-period progression may also apply to C-Mira, though more work is required on this matter.

Miras are becoming of increasing interest as distance indicators since, in the near infrared, they are the brightest objects of the old and intermediate age populations. For LMC O-Miras, Whitelock et al. (2008) found,

$$M_K = -3.51(\pm 0.20)(\log P - 2.38) + 11.241(\pm 0.026) - \text{Mod(LMC)} \quad (16.18)$$



\blacksquare Fig. 16-6

The $M_K - \log P$ relations for LMC O-Miras (solid symbols and dashed line) and C-Mira (open symbols and dotted line) (Reproduced from Whitelock et al. (2008) MN, 386, 313)

⁹Full light curves of Miras at wavelengths longer than $\sim 3.5 \mu\text{m}$ are rare, and bolometric magnitudes are often derived using bolometric corrections to K , based on colors. Such magnitudes may contain systematic errors but are valuable as distance indicators so long as they are derived consistently.

and for C-Miras,

$$M_K = -3.52(\pm 0.36)(\log P - 2.38) + 11.149(\pm 0.047) - \text{Mod(LMC)} \quad (16.19)$$

Thus, there is no observed difference in the slopes. The C-Miras may be slightly brighter than the O-Miras at the same period.

Ita et al. (2004) found $\alpha = -3.59 \pm 0.06$ in \blacklozenge Eq. 16.17 for a large sample of LMC variables, most of which are likely to be Miras, with periods from the OGLE survey (Udalski et al. 1997) and K_s observations. These stars were selected by $J - K_s$ color to be mainly O-Miras. In the SgrI field of the Galactic Bulge, Glass et al. (1995) found $\alpha = -3.55 \pm 0.35$. The larger uncertainty in this case is due to the finite depth of the Bulge. In the giant elliptical Cen A (NGC 5128), Rejkuba (2004) found $\alpha = -3.37 \pm 0.11$ for a sample selected by color to be mainly O-Miras. These results suggest that there is little, if any, variation in PL slope from one environment to another.

At mid-infrared wavelengths Miras are as bright or brighter than Cepheids. For instance, at 8 microns, a 380 day Mira is ~ 1 mag brighter than a 50-day Cepheid. They are thus likely to become important extragalactic distance indicators in the JWST era.

2.4.1 Zero Point for O-Miras

1. Adopting $\alpha = -3.51$ in \blacklozenge Eq. 16.17, Whitelock et al. (2008) found $\beta = -7.32 \pm 0.10$ using revised Hipparcos parallaxes.

A potential problem with deriving trigonometrical parallaxes of Miras is their large physical size (radii in some cases of the order of an astronomical unit). This means that the angular diameters are comparable with the parallaxes. Thus, the parallaxes of the subset discussed by van Leeuwen et al. (1997) range from 2 to 10 mas, while, the angular diameters of those that have been measured by interferometry or lunar occultations range from 17 to 34 mas with the measured diameter, depending strongly on the wavelength due to the deep, complex, stellar atmosphere (see Table 3 of Whitelock and Feast 2000). It is also known that Miras do not present uniform, circular discs (e.g., Karovska et al. 1991). Observations of R Leo with the HST fine guidance sensor in two orthogonal directions show that one axis exceeds the other by 11% and even bigger differences are found in some stars (Lattanzi et al. 1997). The cause of this is not properly understood. It might be non-radial pulsation, or rotation, or it might be star spots resulting from large convection cells. Clearly, if the disc structure changes on a time scale comparable with the period over which the parallax is measured, this could seriously affect the derived distance of an individual Mira. However, one expects these effects to be random and to average out in the mean over a group of stars.

2. Radio frequency methods are becoming of increased importance in deriving Mira distances. The first to be introduced was the method of radio phase-lags. OH masers are detected from the tenuous outer shell of some Miras. The maser lines are typically double peaked; the peaks coming from the front and back of the shell. Since the masers are stimulated by the stellar radiation, they vary in the pulsation period of the star. There is then a phase lag between the two peaks due to the light travel time across the shell whose diameter in the line of sight can thus be measured and the distance deduced by comparing this with a measured angular diameter of the shell and assuming spherical symmetry (see van Langevelde et al. 1990). Strong OH masers are mainly found in very long period Miras (OH/IR stars) with thick dust shells where there is strong circumstellar absorption at K. They are thus not at present

helpful in calibrating a PL(K) relation. Their estimated M_{bol} values (Whitelock et al. 1991) are consistent with an extrapolation of a $PL(M_{bol})$ from shorter periods. Individually, the phase-lag results are not as accurate as VLBI parallaxes.

3. Very long baseline interferometry (VLBI) of OH maser sources can achieve accuracies in position, proper motion, and parallax comparable to the best achieved by Hipparcos in the optical for bright stars. To do this, a fixed standard needs to be within a few degrees for phase-referencing purposes. Extragalactic radio sources are used for this. VLBI OH parallaxes have been published for five Miras by Vlemmings et al. (2003) and Vlemmings and van Langeveldt (2007). From these, Whitelock et al. (2008), find $\beta = -7.08 \pm 0.17$.
4. Trigonometrical parallaxes of Miras can also be measured using VLBI observations of H₂O masers in the Mira shell. For instance, a preliminary parallax of the Mira T Lep measured with the Japanese VERA network (Nakagawa et al. 2009) is 2.62 ± 0.33 mas showing the high accuracy that can be achieved. This yields $\beta = -7.17 \pm 0.28$.
5. Some metal-rich globular clusters contain Miras, and these can be used for calibration with assumed cluster distances. In this way, one obtains (Whitelock et al. 2008) $\beta = -7.34 \pm 0.13$.

This is a secondary method, relying on a calibration of the cluster distances (see [▶ Sect. 2.13](#)). There are a number of uncertainties with this method. The Miras are found in metal-rich clusters, and several of these have high, and sometimes uncertain reddenings. Distance in most cases are based on an adopted absolute magnitude of the horizontal branch. In the case of the key cluster 47 Tuc, the modulus, 13.38, is derived from main sequence fitting Carretta et al. 2000). Unfortunately, there are still some remaining questions regarding this modulus (see [▶ Sect. 2.13](#)).

Lacour et al. (2009) have demonstrated in the case of the Mira χ Cyg that it is possible to obtain a pulsation parallax using infrared interferometry to measure angular radius variations together with a velocity curve from CO infrared lines.

In the later discussions, a value of $\beta = -7.25 \pm 0.07$ (Whitelock et al. 2008) will be adopted.

6. There are a number of possible complications in using a PL relation to derive distances of Miras. There are some Mira-like objects which lie above the PL. They are likely to be overtone pulsators (analogous to Cepheid overtone). The evidence suggests that they can be recognized by their light curves (Glass and Lloyd Evans 2003; Rejkuba et al. 2003; Whitelock et al. 2003).

There are a number of sequences (Wood 2000) of low-amplitude stars pulsating in various overtones which are nearly parallel to the Mira sequence at K . The use of these low-amplitude stars as distance indicators is less straightforward since in general it is not known to which overtone sequence a particular field star belongs.

In the heavily obscured region near the Galactic Center, a number of Mira-like stars which have been observed in the mid-infrared seem to be fainter than the $PL(M_{bol})$ relation (Blommaert et al. 1998). Their status remains uncertain.

2.4.2 Zero Point for C-Miras

The revised Hipparcos parallaxes of C-Miras lead to a zero-point for [▶ Eq. 16.17](#) of $\beta = -7.18 \pm 0.37$, adopting $\alpha = -3.52$ (Whitelock et al. 2008). The uncertainty is too large for this to be very useful, and it is probably best, at present, to rely on [▶ Eq. 16.19](#) and an adopted LMC distance.

Whitelock et al. (2009) derive a bolometric relation for C-Miras in the LMC. On the assumption that the modulus of the LMC is 18.50 (see [▶ Sect. 4.2](#)), this can be written;

$$M_{bol} = -3.31(\pm 0.24)(\log P - 2.5) - 4.38(\pm 0.03(int)) \quad (16.20)$$

The radial velocities of C-Miras in our Galaxy show the effects of differential Galactic rotation (Feast et al. 2006), and this together with the value of the Oort constant, A , discussed in Sect. 1.4 leads to a luminosity scale which is similar to that of Eq. 16.20 at $\log P \sim 2.5$ but has a zero-point uncertainty of ± 0.24 .

2.5 δ Sct and SX Phe Variables

The δ Sct and SX Phe stars (sometimes called dwarf Cepheids) are short period ($P < 0.25$ days) variables. They often show complex light variations due to the excitation of both radial and non-radial modes. High-amplitude δ Sct variables (HADS stars) are thought to pulsate mainly in the fundamental radial mode. A useful summary of our understanding of these stars and their potential as distance indicators is given by Poretti et al. (2008). They occur in globular clusters in the same part of the HR diagram as blue stragglers. Thus, some of them may be merged binaries. The ones in globular clusters, and also in dwarf spheroidal galaxies, as well as metal-poor variables in the Galactic field are generally termed SX Phe stars. However, this term is somewhat confusing since, while SX Phe itself is a metal-poor halo-type object and is presumably related to the globular cluster variables, it is likely that a majority of these stars in a dwarf spheroidal like Fornax are higher mass stars evolving from the main sequence, as are the metal-rich variables in the Galactic field which are generally called δ Sct variables.

McNamara et al. (2007) quote a relation for δ Sct stars,

$$M_V = -2.90 \log P - 0.19[\text{Fe}/\text{H}] - 1.27, \sigma = 0.16 \quad (16.21)$$

based on Hipparcos parallaxes, though full details have not been published.

The complications which are liable to arise in practice are demonstrated by the SX Phe stars in the Fornax dwarf spheroidal galaxy. These show a large scatter about a PL relation (Poretti et al. 2008), and the authors suggest that this is due to three overlapping populations: (1) Intermediate age stars evolving off the main sequence through the instability strip and pulsating in the fundamental mode; (2) similar stars pulsating in the first overtone and which are therefore brighter at a given observed period than the fundamental pulsators; and (3) old, blue straggler-type, objects which have lower luminosity due to lower metallicity and/or stellar structure.

It will be clear from the above that, though these stars have considerable potential as distance indicators, much work remains to divide them into recognizable, homogeneous, groups.

2.6 Red Clump

The red giant clump, consisting of core helium burning stars, is found in old open clusters and metal-rich globular clusters. It is the equivalent of the horizontal branch in old metal-poor populations. Paczyński and Stanek (1998) pointed out that stars of this type were plentiful in the solar neighbourhood and, since many of them had accurate parallaxes from Hipparcos, they were potentially good distance indicators. The latest calibration of clump absolute magnitudes from local stars in the I band is that of Groenewegen (2008) based on revised Hipparcos parallaxes (van Leeuwen 2007). Groenewegen also discusses the effects of selection bias in this

■ **Table 16-2**

Theoretical population corrections to the absolute magnitudes of the red clump (Salaris and Girardi 2002)

System	ΔM_V	ΔM_I	ΔM_K
Solar neighborhood	0	0	0
Baade-window (scaled-solar)	-0.21	-0.08	-0.07
Baade-window (α -enhanced)	-0.06	-0.01	-0.11
Carina dSph	+0.59	+0.35	-0.17
LMC	+0.26	+0.20	-0.03
SMC	+0.31	+0.29	-0.07

calibration. He found for the local sample $M_I = -0.22 \pm 0.03$. Similarly, but using extensive new photometry, Laney et al. (2011) have derived $M_{K_s} = -1.61 \pm 0.02$ and redder infrared colours than previously adopted.

There are a number of problems with the use of clump stars to determine distances. The most important are probably the effects of age and metallicity on the absolute magnitudes. These effects were estimated theoretically by Girardi and Salaris (2001) and Salaris and Girardi (2002).

● **Table 16-2** gives their results for the Baade window of the Galactic Bulge and a number of other systems compared with the solar neighbourhood. It will be seen that these corrections can be large and depend critically on the population mix assumed. For instance, Grocholski et al. (2007) who discuss the distribution in depth of old clusters in LMC use the corrections of Salaris and Girardi at K , and these lead to their adopting clump absolute magnitudes ranging from -1.10 to -1.57 for different clusters. In view of the need for such corrections, the red clump scale cannot be considered entirely empirical. However, since in some ranges of age, the metallicity effects are predicted to be of opposite signs in I and K , it may be possible to use this to constrain derived moduli.

In the LMC field, there is a ~ 0.4 mag extension of the red clump to fainter magnitudes and similar results have been found in the intermediate age SMC cluster NGC419 (Girardi et al. 2009). Theory predicts that the main clump is a feature in the evolution of all intermediate age and old stars, whereas the lower secondary clump is populated by stars just massive enough to start core-helium burning in nondegenerate conditions (Girardi 1999).

2.7 Tip of Red Giant Branch

The tip of the red giant branch is a prominent feature of many groups of stars and galaxies. It represents the onset of core helium burning. It has been increasingly used as a distance indicator both in the optical (generally at I) and in the near infrared at K . Considerable care is necessary in deriving a value for the apparent magnitude of the tip, especially when the tip region of a color-magnitude diagram is sparsely populated or in the presence of scatter due to photometric errors. Madore and Freedman (1995) do numerical experiments to show that at least 100 stars are needed within a magnitude of the tip for it to be determined satisfactorily. A variety of

methods have been proposed to obtain unbiased tip estimates (see, for instance, the references in Salaris and Girardi 2005). These methods do not necessarily always yield the same results. For instance, Pietrzyński et al. (2009a) attribute the difference of 0.14 mag in the apparent K magnitude of the tip in the Fornax dwarf spheroidal which they derive from that obtained by Gullieuszik et al. (2007) to the different techniques used. They used a Sobel filter technique (Sakai et al. 1996), whereas Gullieuszik et al. (2007) using a maximum likelihood algorithm (Makarov et al. 2006). Fornax illustrates a further complication. The TRGB has a slope, being brighter on the red side due to a mix of stellar populations (Whitlock et al. 2009).

The calibration of the TRGB absolute magnitudes has until recently depended either on theoretical models or on observations of the tip in systems whose distances are known in some other way. A general review of the theory of the red giant branch is given by Salaris et al. (2002). This includes discussion of both the TRGB and the red clump. Theoretical estimates of TRGB absolute magnitudes in both in I and K are given by Salaris and Girardi (2005). It is worth noting that these authors write,

Our results clearly show that the presence of a well developed RGB in the Color-Magnitude Diagram of a stellar system with a complex SFR does not guarantee that it is populated by globular cluster-like red giants, and therefore the TRGB method for distance determination has to be applied with caution. A definitive assessment of the appropriate corrections for population effects on TRGB distances has however to wait for a substantial reduction in the uncertainties on the BC_I scale for cold stars.

Empirical calibrations have generally been done adopting distances (e.g., of globular clusters) derived using assumed absolute magnitudes for HB or RR Lyrae or red clump stars. A useful review of calibration methods has been given by Bellazzini (2008). Perhaps the most complete analysis of the calibration is given by Rizzi et al. (2007a). They compile TRGB data for a number of galaxies, taking $(V - I)_o$ as a proxy for metallicity. This leads to a value of the slope b in the equation:

$$M_I^{TRGB} = a + b(V - I)_o. \quad (16.22)$$

They then fix the zero point, a , from observations of a few local group galaxies whose distances are derived from assumed HB absolute magnitudes. This is 0.11 mag fainter than that derived in [Sect. 2.3](#) and, for consistency, their relation was changed to this scale.

$$M_I^{TRGB} = -4.16(\pm 0.04) + 0.22(\pm 0.01)[(V - I)_o - 1.6] \quad (16.23)$$

The errors are their internal values.

Bellazzini (2008) derived a slightly different relation. The basic distance scale used in that work is fixed by the distance to the cluster ω Cen derived from the double-lined eclipsing variable it contains. Unfortunately, the error on this is still substantial (± 0.11 mag).

Ferraro et al. (2000) obtained a relation for the TRGB in K_s also from globular clusters and based on a ZAHB scale. Converting this to the scale of [Sect. 2.3](#) leads to the relation:

$$M_K = -7.01(\pm 0.14) - 0.64(\pm 0.12)[M/H] \quad (16.24)$$

There are evidently still considerable uncertainties in the absolute calibration of TRGB distances just discussed. Bellazzini (2008) comments that “the zero point is known with uncertainties of ± 0.12 in the best case; TRGB distance moduli with error bars smaller than this figure neglect part of the actual error budget.”

A recent calibration of the TRGB in the solar neighbourhood at K has been carried out by Tabur et al. (2009) using revised Hipparcos parallaxes. Their result depends on a sample containing stars with substantial percentage errors in their parallaxes ($\sigma_\omega/\omega \leq 0.1$ for the smallest sample to ≤ 0.25 for the largest). The authors discuss standard bias problems of the Lutz–Kelker type. (see [Sects. A.4](#) and [A.5](#)). However, the problem is nonstandard in that what is sought is the absolute magnitude of the edge of a distribution rather than a mean value. Qualitatively, one might expect that the stars placed at the TRGB would tend to have underestimated rather than overestimated parallaxes, leading to an apparent increase in TRGB brightness with increasing relative parallax error. Tabur et al. do in fact find a small effect of this kind (which is of opposite sign to the standard Lutz–Kelker-type effect). The value the authors adopt for the TRGB is $K = -6.85 \pm 0.03$ (without any bias correction) for a presumed near solar metallicity population. If the trend with parallax error just discussed is real, then the best value is slightly fainter than this. At present, this is the only result based directly on parallaxes. Tabur et al. were unable to get a satisfactory result in I from parallaxes. In the following, the zero point has been kept as in [Eq. 16.24](#) so that the M_I and M_K scales are the same.

2.8 Novae

The use of novae as distance indicators was reviewed by Gilmozzi and Della Valle (2003). The classic work by Arp (1956) on M31 novae clearly established that there was a relation between absolute magnitude at maximum and the rate of decline. A Galactic calibration (Downes and Duerbeck 2000) is based on expansion rates of nova shells. Apparent differences from a relation based on M31/LMC novae may be due to selection effects. Because of their brightness at maximum ($M_V \sim -9$ for those with the fastest declines), novae are, in principle, attractive distance indicators, but they will require more work before they are really useful.

2.9 Eclipsers

As mentioned in [Sect. 1.3](#), determining distances to eclipsing variables has some similarity to the pulsation parallax method. The diameters of the stars are determined by combining light and radial velocity curves. The surface brightness (i.e., the effective temperature) of the primary is estimated from either an empirical calibration based on color or by matching observed spectra with theoretical ones. The method is sensitive to the reddening which is either adopted from external evidence or deduced from a comparison of spectrophotometry and theoretical models. A useful review of eclipsing binaries as standard candles is given by Clausen (2004).

2.10 Spectroscopic Parallaxes

Much of the work on the distances and distributions of common stars relies on the estimation of spectroscopic parallaxes or its extension, the use of narrow-band photometry. The calibration of these systems is a two-step process; selecting luminosity sensitive line (or narrow band) strengths or ratios and calibrating them in some way, e.g., using measured trigonometrical parallaxes, cluster membership, etc. The early work in this field is reviewed by Blaauw (1963), who discusses in some detail the bias problems that can arise in the calibration process. This section

discusses the calibration of the MK spectral-type system as well as the special problems of the OB stars and supergiants. The use of narrow-band indices is discussed in [Sect. 2.11](#).

2.10.1 The MK System

The MK system of spectral type and luminosity classification (Morgan et al. 1943; Morgan and Keenan 1973) has been extensively used in the past for distance estimation. A calibration of the system in absolute magnitudes was given by Blaauw (1963) and updated by Schmidt-Kaler (1982), whose results have been widely used. For main sequence stars, apart from OB types, and for later-type giants, the calibrations were primarily based on trigonometrical and statistical parallaxes. For the OB stars and supergiants, data from open clusters were used with distances derived from main sequence fitting (see [Sect. 2.12](#)). The methods are discussed in detail by Blaauw (1963).

It will be apparent from the Appendix that in the use of any calibration, it is important to know not only the mean absolute magnitude of a given class of stars, but also the intrinsic dispersion about the mean, and it is not straightforward to derive this from a magnitude selected sample. It is also important to know whether one is dealing with an absolute magnitude per unit volume or for a selection by apparent magnitude.

The data in the Hipparcos catalogue and its revision (van Leeuwen 2007), especially the new parallaxes, should provide the basis for a revision and improvement of the calibration. Perryman (2009) briefly summarizes a number of papers that deal in part with this problem. However, in some of them, it is not clear that the method of analysis is statistically sound. The only worker who has attempted a complete revision of the MK luminosity scale based on Hipparcos parallaxes is (Wegner 2006, 2007) and the discussion in [Sect. A.5](#) suggests that further analysis is desirable.

2.10.2 OB Stars and Supergiants

Distance scales for OB stars (stars of type earlier than B5 of all luminosity classes) and later-type supergiants are of special importance. Their high luminosities allow them to be studied to large distances in our own Galaxy as well as in nearby galaxies, and they are traces of a young stellar population. Spectroscopic parallaxes of supergiants gave direct evidence for spiral structure in our own Galaxy (Morgan et al. 1953).

The importance of bias corrections in using OB stars is discussed in [Sects. A.2](#) and [A.3](#). As mentioned in the last section, the luminosity calibration of these stars has rested primarily on young, open, star clusters. The general method is to start with a cluster of known distance and to derive distances to other clusters by a main sequence fitting procedure. The problem of cluster distances is discussed in [Sect. 2.12](#). Blaauw (1963) begins with the Hyades main sequence and fits to it younger and more distance clusters. A problem with this method is that errors accumulate in the process. Thus, the standard error in the distance modulus of η and χ Persei, which is an important source of absolute magnitudes of OB stars and supergiants, is between 0.3 and 0.4 mag. A ZAMS from Schmidt-Kaler (1982) is sometimes used for cluster distances, but a study of nearby stellar association based on Hipparcos parallaxes and proper motions suggests that this is too bright by about 0.2 mag (e.g., Brown et al. 2000).

For early-type stars, an alternative to the MK system is to use the strengths of Balmer lines as a luminosity criterion. Both the equivalent width of H γ (see Balona and Crampton 1974) and the measurement of H β in the Strömgen narrow band photometric system have been used. Again, calibration rests on clusters. The calibration for OB stars in the H β system depends on the Pleiades (Crawford 1978) as a starting point.

As with the MK system in general, the calibration of OB stars and supergiants would be greatly strengthened by a comprehensive analysis of Hipparcos data.

2.10.3 The Most Luminous Supergiants

Blue and red supergiants of the highest luminosity have for a long while attracted attention as possible distance indicators for systems containing a young, massive, population. In recent times, there has been progress in establishing them as useful distance markers. In the case of the blue supergiants, useful reviews of their potential as distance indicators have been given by Kudritzki and Przybilla (2003) and by Bresolin (2003). The method used for these stars essentially depends on the comparison of spectra with NLTE models. Two methods of determining luminosities have been developed. The first is a relation between the momentum of the wind $\dot{M}v_{\text{inf}}$ which is being driven from the star by the, luminosity (L)-dependent radiation pressure. Both \dot{M} and v_{inf} can be derived from spectra. The second is a flux-weighted gravity-luminosity relation which relies only on determination of surface gravity and effective temperature. Such a relation is expected, at least approximately, since supergiants evolve in the HR diagram at nearly constant mass and luminosity. Then, with M the mass of the star, R its radius, g the surface gravity, and T_{eff} the effective temperature, and since $L \propto R^2 T_{\text{eff}}^4$, $M \propto gR^2$, and for these stars L is a function of M , it follows that L is a function of g/T_{eff}^4 . The calibration of these relations depends at present on luminous supergiants in local group galaxies for which distances have been derived by other means.

The red supergiant variables also have considerable potential as distance indicators. In the Magellanic Clouds, such stars were studied by Wood et al. (1983) and others, in M33 by Kinman et al. (1987) and in M101 by Jurcevic et al. (2000). However, more work is required and at present an absolute magnitude calibration must depend on other distance indicators (e.g., a distance to the LMC).

2.11 Large-Scale Studies of Common Stars

The most extensive study in recent times of F and G dwarfs in the solar neighborhood has been the Geneva–Copenhagen survey which is based on Strömgen $uvby\beta$ narrow band photometry, radial velocities, Hipparcos parallaxes, and Tycho-2 proper motions. From these data, absolute magnitudes, metallicities, temperatures, and kinematics have been obtained for $\sim 14,000$ stars (Holmberg et al. 2007). Hipparcos parallaxes were used to calibrate absolute magnitudes in terms of a (rather complex) relation which uses the measured narrow-band magnitudes.

Another similar relation for the metallicities is calibrated using results of high-resolution spectroscopy. The colors are used to give the temperatures. An important result from this work is an age-velocity dispersion relation. Ages of individual stars are estimated by comparing their position in an HR diagram with theoretical evolutionary tracks.

Attempts to understand the structure of our Galaxy using star counts has a long history. Such studies have been transformed in recent times by the ability to obtain multicolor observations of good accuracy to faint limits over large areas. Two such studies have been that of Siegel et al. (2002) who summarize earlier surveys and Juric et al. (2008) who analyze data from the Sloan Digital Sky Survey (SDSS). The latter paper analyzed the distribution of ~ 48 million Galactic stars out to distances of ~ 20 kpc. Both these studies are restricted to relatively cool stars. In that case, most of the stars are expected to be on the main sequence, and the effect of giants and subgiants can either be ignored as minimal or corrected for. The distance-scale problem then reduces to a calibration of the main sequence absolute magnitudes as a function of color and their dependence on metallicity. Siegel et al. obtain a solar metallicity main sequence using Hipparcos parallaxes and estimate metallicity effects from Hipparcos subdwarf data. Juric et al. use a main sequence from the globular cluster M13 and a version of the method of statistical parallaxes and do not explicitly account for metallicity. As these authors show, there is considerable uncertainty at the faint end of the main sequence in their work. As in the globular cluster work ([▶ Sect. 2.13](#)), this is mainly connected to metallicity effects. These are further discussed in the SDSS system by Klement et al. (2009).

It is expected that in the near future the number of stars with accurately measured trigonometrical parallaxes will greatly increase. A number of space-based projects have been proposed (see, e.g., Perryman 2009). Two of the major projects are the space astrometry satellite, GAIA (see, e.g., Perryman 2003, and GAIA website) and the proposed ground-based Large Synoptic Survey telescope (LSST) (e.g., Ivezić et al. 2008). GAIA is expected to measure the parallaxes of 15 mag stars with a standard error of about $20\mu\text{as}$ (Lindegren et al. 2008). The LSST will obtain parallaxes of very faint stars. The standard error of a parallax is expected to be ~ 1.3 mas at 23 mag. These projects therefore hold out the prospect of studying the solar neighbourhood and the wider Galaxy in detail down to very faint limits based directly on individual trigonometrical parallaxes.

2.12 Open Clusters

A knowledge of the distances to open clusters and associations in our Galaxy is important for comparison with stellar evolutionary models, for a study their spatial distribution as a function of age and metallicity, for their relation to the field stars in which they are imbedded, and for estimating the luminosities of their members (e.g., Cepheids and supergiants see [▶ Sects. 2.1](#) and [▶ Sect. 2.10.2](#)). The distance of the Hyades and the moving group of which it is a part have been well determined from Hipparcos parallaxes and proper motions (see the summary by Perryman (2009)). The mean distance modulus of the cluster itself is 3.33 ± 0.01 (Perryman et al. 1998). Parallaxes and proper motions have been used to determine the distances of the nearer stellar associations (de Zeeuw et al. 1999). For distant clusters, the distances are normally derived by main sequence fitting. (see, e.g., An et al. 2007). Considerable care is needed in this procedure. Because of the steepness of the main sequence, the resulting distance is very sensitive to the adopted reddening. Furthermore, the main sequence is likely to be widened by binaries. The template ZAMS used has generally been either that of the Pleiades whose distance is presumed known or a theoretical sequence. An et al., for instance, adopt a Pleiades distance modulus of 5.63 ± 0.02 . This gives a ZAMS in agreement with predictions of stellar evolution models, and distances close to this are suggested by some other methods (binary stars, an HST parallax, etc.). However, a smaller modulus (5.40 ± 0.03) was determined by van Leeuwen (2009)

from revised Hipparcos parallaxes. The difference between these values is not currently understood (see summaries and discussions in Perryman (2009) and van Leeuwen (2009) and, so far as it impinges on the Cepheid scale through the distances of clusters containing Cepheids, Feast 2003).

2.13 Globular Clusters

A main interest in establishing a reliable distance scale for globular clusters is to determine their relative and absolute ages as well as defining the large-scale structure of the Galaxy as a function of metallicity and age. Absolute ages give a lower limit to the age of the Universe, and relative ages can establish whether there is a dependence of metallicity on age. Ages are derived from main sequence turnoffs, and a 0.1 mag. error in distance modulus leads to an error of about 1 Gyr in age. Cluster distances can be obtained from main sequence fitting methods. This became feasible when Hipparcos provided good parallaxes of nearby subdwarfs. A useful summary with references to earlier work is given by Gratton et al. (2003).

Bias corrections to the Hipparcos data are very small, and the main uncertainties come from photometric calibration, reddening uncertainty, and metallicity scale. The best data give distance moduli of globulars with a s.e. of ~ 0.08 mag. However, there are remaining uncertainties. These are illustrated by conflicting results on the key metal-rich cluster 47 Tuc which may be partly due the different adopted metallicities for the cluster (see Gratton et al. (2003)). Problems in the use of a white dwarf cooling sequence to determine distances to globular clusters are discussed by Salaris et al. (2001).

In deriving cluster ages, there is the further uncertainty in the models, including the uncertain effects of microscopic diffusion and the realization, subsequent to the work of Gratton et al. that a laboratory revision of the rate of the important reaction, $^{14}\text{N}(p,\gamma)^{15}\text{O}$ increased estimated ages by ~ 0.8 Gyr (Degl'Innocenti et al. 2004; Imbriani et al. 2004).

Benedict et al. (2011) give revised distances and ages of a number of globular clusters based on the RR Lyrae scale discussed in [Sect. 2.3](#).

3 The Scale of Our Galaxy

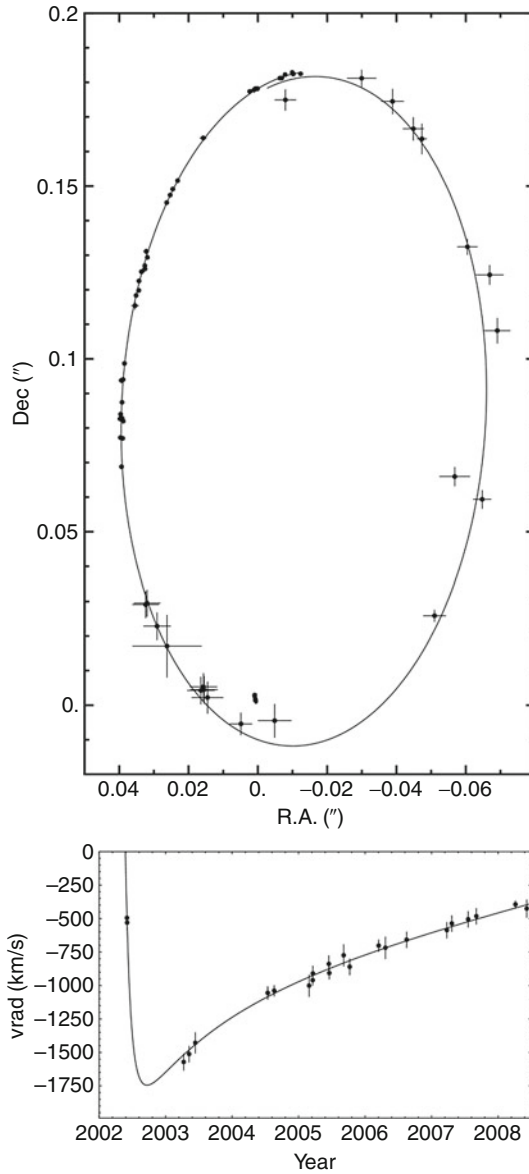
3.1 Distance to the Galactic Center

A review of methods for determining the distance to the Galactic Center R_0 was given by Kerr and Lynden Bell (1986). They suggested that the best mean value at that time was 8.5 ± 1.1 kpc. The sections below concentrate on more recent work.

3.1.1 Orbits Around the Central Black Hole

One of the most remarkable achievements of modern astronomy has been the ability to monitor, astrometrically, stars close to the central blackhole of our Galaxy (within ~ 0.3 arcsecs or 0.01 pc), to define their orbits, and also to measure their orbital radial velocities. These stars have short orbital times, ~ 16 years in one case and very high orbital velocities ($\sim 1,000$ km s $^{-1}$) (see [Fig. 16-7](#)). From this work come estimates of the mass of the blackhole and its distance.

Two groups have been particularly active in this work (e.g., Ghez et al. 2008; Gillessen et al. 2009a). The latter group (Gillessen et al. 2009b) obtained $R_0 = 8.28 \pm 0.33$ kpc.



■ Fig. 16-7

The orbit and radial velocity curve of star S2 round the Galactic Centre (Gillessen et al. 2009a, reproduced by permission of the AAS)

3.1.2 Miras

Glass and Feast (1982) showed that Miras in the Galactic Bulge can be used to estimate a distance to the Galactic Center. The problem is somewhat complicated by the fact that the Miras

there probably belong to a bar-like structure (e.g., Whitelock and Catchpole 1992) and corrections are needed if the fields studied are offset from the direction to the Center. This problem is avoided in the work of Matsunaga et al. (2009b) who made JHK_s observations of Miras in an area of 20×30 arcmins in the direction of the Center. The region is heavily reddened, and corrections for this were estimated using period-color relations. The filter system used is slightly different to that used in the absolute magnitude calibration (▶ Sect. 2.4). However, the results of Matsunaga et al. can be reduced to the same scale through a comparison of LMC PL relations. Then on the scale adopted in ▶ Sect. 2.4, the Matsunaga et al. result becomes 8.39 ± 0.54 kpc.

3.1.3 Red Clump

The results at K_s of Nishiyama et al. (2006) together with the results of Laney et al. (2011) ▶ Sect. 2.6 give 7.87 ± 0.34 kpc. This includes a population correction from Salaris and Girardi (2002) of -0.07 mag. A factor 0.98 smaller in distance would be obtained if the correction for an enhancement of α elements (as expected in the Bulge, e.g., Rich and Origlia 2005) had been used.

3.1.4 Type II Cepheids

The calibration of the infrared PL relation in ▶ Sect. 2.2 together with observations of Groenewegen et al. (2008) of type II Cepheids in a region around the Galactic Centre leads to $R_o = 7.63 \pm 0.21$ kpc.

3.1.5 RR Lyraes

There has been a considerable amount of work at optical wavelengths on RR Lyraes in the region of the Galactic Bulge (summarized by Smith (1995)). The uncertainty in this work is due mainly to the question of interstellar absorption. Groenewegen et al. (2008) have obtained K_s observations of RR Lyraes in a region around the Centre. Based on the calibration of ▶ Sect. 2.3, these lead to $R_o = 8.32 \pm 0.62$ kpc.

3.1.6 Kinematic Determination

For stars such as OB stars or Cepheids which are moving on circular or nearly circular orbits, those close to the solar circle will have zero velocity with respect to the circular velocity at the Sun. The identification of such stars at large distances from the Sun leads to an estimate of R_o . For instance, the observed radial velocity V_r , corrected for local solar motion is:

$$V_o = R_o(\omega - \omega_o) \sin l \cos b \quad (16.25)$$

where the symbols are as defined in ▶ Sect. 1.4. For $R - R_o$, small this can be approximated as:

$$V_o = -A(R - R_o) \sin l \cos b \quad (16.26)$$

The Oort constant A is known from proper motions (e.g., [Sect. 1.4](#)), and there is also the geometrical relation:

$$R^2 = R_o^2 + r \cos^2 b - 2R_o r \cos l \cos b \quad (16.27)$$

where r is the distance of the object from the Sun. A first estimate of R_o by this method (Feast and Thackeray 1958) gave $R_o \sim 8.9$ kpc, and there are several later estimates. Pont et al. (1994) used Cepheids radial velocities to determine Galactic constants including R_o . With a slight correction to bring this onto the Cepheid scale of [Sect. 2.1](#), their result becomes 7.9 ± 0.3 kpc.

More recently, Reid et al. (2009a) have extended the method using the space motions and distances of 18 star-forming regions derived from VLBI and VLBA measurements of parallaxes, proper motion, and radial velocities of the masers the regions contain. From these data, they find $R_o = 8.4 \pm 0.6$ kpc.

There are a number of problems associated with this method. At least in the optical work, the objects with high weight in the determination will be at large distances from the Sun and may well be among the most distance in a given sample. These distances may be significantly affected by bias (see [Sect. A.3](#)). In addition, if the objects are in a limited volume of space, systematic deviations from circular velocity can occur. For instance, the peculiar velocities of Cepheids, and also of OB stars, tend to be correlated in regions of the order of 1 kpc size.

The results derived by this method are sensitive to the adopted value of the component, v_o , of the local solar motion in the direction of Galactic rotation. This is easily seen in the case of radial velocities since the correction from the observed radial velocity V_r to V_o contains the term $v_o \sin l \cos b$ and cannot be separated from the rhs of [Eq. 16.26](#).

Unfortunately, there has been some uncertainty as to the correct value of v_o . Feast and Whitelock (1997) found that the kinematics of Cepheids gave $v_o = 11.2 \pm \sim 1.0$ km s⁻¹ in good agreement with many earlier determinations from young objects (e.g., Delhaye 1965). Fitting the mean deviation from circular motion (the asymmetric drift) as a function of velocity dispersion (the Stromberg parabola) and extrapolating to zero dispersion also gives an estimate of v_o . In this way, Dehnen and Binney (1988) found the much smaller value of 5.25 ± 0.62 km s⁻¹ and many authors have adopted this value, though reasons have been given (Feast 2000) for preferring the larger value which depends on observations over a significant volume of the Galaxy. A new analysis by Binney (quoted by McMillan and Binney (2010)) of the asymmetric drift now finds a result in good agreement with the larger value. Adopting this, a reanalysis of the Reid et al. data by McMillan and Binney suggest $R_o = 7.6 \pm 0.5$ kpc.

3.1.7 Parallax of Sgr B2

Sgr B2 is a high-mass star-forming region close to the Galactic Center. Parallax observations of H₂O masers from this region using VLBA have been obtained by Reid et al. (2009b). Sgr B2 is believed to be about 0.13 kpc nearer than the Center, and applying this offset, one obtains $R_o = 7.9 \pm 0.8$ kpc. It is expected that further measurements will refine this result.

3.1.8 Classical Cepheids

Three classical cepheids in the galactic nuclear bulge give $R_o = 7.89 \pm 0.36$ kpc. (Matsunaga et al. 2011).

■ **Table 16-3**

Estimates of R_o , the distance to the Galactic Centre

Method	R_o , kpc	Mod.	Weight
Central orbits	8.28 ± 0.33	14.59 ± 0.09	2
Miras	8.39 ± 0.54	14.62 ± 0.12	1
Clump	7.87 ± 0.34	14.48 ± 0.10	1
Cepheids	7.63 ± 0.21	14.41 ± 0.06	1
RR Lyraes	8.32 ± 0.62	14.60 ± 0.18	1
Kinematics	7.6 ± 0.5	14.40 ± 0.14	1
Sgr B2	7.9 ± 0.8	14.49 ± 0.2	1
Cepheids	7.89 ± 0.36	14.49 ± 0.10	1
Mean	8.0	14.52	

■ **Table 16-4**

Estimates of the Disc scale length

Method	Scale Length	Objects	Ref.
Counts	2.60 ± 0.52 kpc	K/M dwarfs	Juric et al. 2008
Counts	3.02 ± 0.12 kpc	A stars	Sale et al. 2010

3.1.9 Summary

These values are listed in [Table 16-3](#). Proper weighting remains difficult since some of the results may be affected by systematic errors. The listed means gives double weight to the results from orbits round the central black hole. The true uncertainty in mean R_o is difficult to assess but is probably less than 0.5 kpc.

3.2 The Scale Length of the Galactic Disc

If, as suggested by observations of other galaxies, the surface brightness of the Galactic disc, $I(R)$, is exponential, then:

$$I(R) = I_o \exp(-R/R_d) \quad (16.28)$$

where R_d is the scale length of the disc. This quantity is of importance, for instance, in comparing our Galaxy with other spirals. Much of the early work on the determination of R_d is summarized by van der Kruit (2000) who gives reasons for accepting some determinations and questioning others. These early results suggested a scale length of ~ 4.5 – 5 kpc.

More recent data ([Table 16-4](#)) give smaller values.

The first entry is from work using the SDSS survey discussed in [Sect. 2.11](#). The authors discuss the possible effects of statistical bias. The second is from an $H\alpha$ and photometric survey of $\sim 40,000$ A-type stars in the anticentre direction.

It is possible to attempt a determination of R_d from stellar kinematics (Dehnen and Binney 1988; Feast 2000). However, as in [Sect. 3.1.6](#), the result depends on the adopted value of ν_o .

4 Comparison of Distance Scales

To compare the calibrations of the various distance indicators discussed above and also to illustrate some of the problems in distance estimation, this section considers in detail distance estimates for the LMC, the Fornax dwarf spheroidal, and the spiral NGC 4258.

4.1 NGC4258

The spiral galaxy NGC4258 offers a particularly useful opportunity to test the classical Cepheid scale discussed in [Sect. 2.1](#). Herrnstein et al. (1999) have measured the distance to this galaxy from the motions of H_2O masers in orbit around the central black hole. The central masers moving across the line of sight measure the rotation of the maser system in arcsecs/year, while the radial velocities of outer, high-velocity masers measure this in km s^{-1} . Combining these sets of measurements leads to a distance modulus of 29.29 ± 0.15 .¹⁰ Macri et al. (2006) have obtained observations of Cepheids in an inner region of this galaxy where the metallicity is close to that in the solar neighbourhood. Applying [Eq. 16.7](#) to their data yields a distance modulus of 29.22 ± 0.03 . The agreement is evidently good.

Rizzi et al. (2007a) give I and $(V - I)$ for the TRGB. Together with [Eq. 16.23](#) of [Sect. 2.7](#), this gives a modulus of 29.53 ± 0.12 . The uncertainty of the modulus comes from the discussion in [Sect. 2.7](#). The difference from the maser and Cepheid results ($\sim 2\sigma$) is marginally significant.

4.2 LMC

4.2.1 Classical Cepheids

A distance to the LMC may be obtained by comparing PL relations for LMC Cepheids with Galactic Cepheids of known parallax. As discussed in [Sect. 2.1](#), there remains some uncertainty in the slope of the Galactic PL(VI) relation at least for periods greater than 10 days. However, the OGLE survey has given V, I data for a large number of LMC Cepheids with periods shorter than this, and a comparison with HST/Hipparcos parallaxes of Galactic Cepheids in this period range is essentially free of any uncertainty in the slope, α of [Sect. 2.1](#). The observations of LMC Cepheids in this period range at K are less plentiful (Persson et al. 2004), but the overlap with the Galactic trigonometrical parallax calibrators seems quite satisfactory. In this way, one obtains (van Leeuwen et al. 2007) an LMC modulus $(m - M)_0 = 18.52 \pm 0.03$ from PL(VI) and 18.47 ± 0.03 from PL(K). The pulsation parallax results of Fouqué et al. (2007), and the earlier results of Benedict et al. (2007) agree with these values. The slightly greater mean modulus (18.56) obtained by Di Benedetto (2008) is probably due to use of Galactic data at longer period and where it differs from the careful discussion of Fouqué et al. who note the good agreement of trig parallaxes, pulsation parallaxes, and distances derived for Cepheid in open clusters for periods less than 10 days.

The LMC Cepheids are metal poor compared with the Galactic absolute magnitude calibrators. The results given above do not take into account any dependence of the absolute magnitudes on metallicity. If the metallicity correction of Macri et al. (2006) were adopted,

¹⁰According to Riess et al. (2009), unpublished data reduce this standard error to ~ 0.07 .

the LMC Cepheid modulus in VI would be 18.39. However, the result discussed below for the RR Lyrae variables, which agrees well with an uncorrected Cepheid value, suggests that any Cepheid metallicity correction is likely to be small in VI and K and the uncorrected value is used here.¹¹

4.2.2 Type II Cepheids

Using the calibration of the PL relations in $W(VI)$ and K discussed in [Sect. 2.2](#), the data of Matsunaga et al. (2009a) lead to the LMC moduli given in [Table 16-5](#).

4.2.3 RR Lyraes

Borissova et al. (2009) discussed the distance of the LMC based on K band observations of RR Lyrae variables. Their adopted absolute magnitude calibration (their [Eq. 16.12](#)) is identical to [Eq. 16.15](#) above, though the latter has a much smaller zero-point error. Thus, the RR Lyrae modulus of the LMC is 18.53 ± 0.04 . The mean metallicity of the LMC sample is close to that of the Galactic calibrators. Thus, any uncertainty in the metallicity term has a negligible effect on the result. The agreement of this modulus with the Cepheid modulus uncorrected for

Table 16-5
Distance Modulus of the LMC

Method	Mod
Cepheids (VI)	18.52 ± 0.03
Cepheids (K)	18.47 ± 0.03
RR Lyraes (K)	$18.53 \pm \sim 0.04$
CephII (VI)	$18.46 \pm \sim 0.1$
CephII (K)	$18.50 \pm \sim 0.1$
Eclipsers (OB)	~ 18.4
Eclipser (G)	$18.50 \pm (0.06)$
Miras	18.49 ± 0.07
SN Ring	$18.56 \pm (0.05)$
SN model	$18.5 \pm (0.2)$
Clump	18.47 ± 0.02
TRGB (VI)	$18.68 \pm \sim 0.1$
TRGB (K)	$18.50 \pm \sim 0.1$
δ Sct	$18.48 \pm \sim 0.1$
Mean	18.50

Standard errors of single objects in brackets

¹¹ Feast and Catchpole (1997) obtained 18.70 ± 0.10 for the LMC modulus from the initial release of Hipparcos parallaxes and B, V data of Galactic and LMC Cepheids. The difference from the present value is primarily due to the improvement in LMC data (mainly due to OGLE) and the possibility of using V, I photometry. The new trigonometrical parallaxes, although more accurate than those used in 1997, are not systematically different.

metallicity effects suggests (as just noted) that any such effects on the Cepheids is small. It has however to be realized that the structure of the LMC in depth becomes of importance when distance moduli from different distance indicators are being compared at the $\sim \leq 0.1$ mag level. The result of Borissova et al., for instance, depends on the adoption of a model for the structure of the LMC.

It is possible to obtain an LMC modulus using the RR Lyrae $M_V - [\text{Fe}/\text{H}]$ relation (Eq. 16.16). In this way, the results of Gratton et al. (2004) lead to 18.61 ± 0.08 and the OGLE data (Soszyński et al. 2003) to 18.44 ± 0.06 . The difference between these values is primarily due to the different values of the interstellar absorption adopted, and since this matter is not entirely settled, it seems best at present to rely on the K result which is relatively insensitive to the adopted interstellar absorption.

4.2.4 Eclipsers

Most of the eclipsing systems in the LMC which have been studied so far are of early type (OB stars). The principal uncertainty in that case is probably the derivation of the surface brightness, either from colors or spectrophotometry (Sect. 2.9). These methods are sensitive to the interstellar reddening adopted, and this may account for some of the spread in values found by different observers for the same star. There are also considerable differences between distances for different stars (e.g., HV 932. $\text{Mod} = 18.63 \pm 0.08$ (Clausen et al. 2003) and HV5936, $\text{Mod} = 18.18 \pm 0.08$ (Fitz-Patrick et al. 2003), an apparent depth spread of 10.0 ± 2.6 kpc). Such a large depth was unexpected for a young population. The mean of these extreme values is near 18.4. Recently, Pietrzyński et al. (2009b) have derived a modulus of 18.50 ± 0.06 from a LMC eclipsing system containing two bright G giants (luminosity class II according to the Schmidt-Kaler calibration).

4.2.5 Miras

The PL(K) zero-point (-7.25) (Sect. 2.4.1) and LMC data discussed there leads to an LMC modulus of 18.49 ± 0.07 .

4.2.6 SN1987A

The 1987 type II supernova in the LMC is surrounded by a circumstellar ring (1.66 arcsec diameter) inclined to the line-of-sight, which was ionized by radiation from the supernova explosion. The measured integrated flux from emission lines rises as ionizing radiation from the supernova explosion lights up, first the part of the ring nearest to us and then more and more of the ring. It then fades as the supernova fades. This time variation, together with the measured angular diameter of the ring, leads to a distance. A small correction is then made, based on a model of the LMC, for the displacement of the supernova from the LMC center, resulting in an LMC modulus of 18.56 ± 0.05 (see Panagia et al. 1991; Panagia 2005). Modelling the spectra gives 18.5 ± 0.2 (Mitchell et al. 2002).

4.2.7 Red Clump

Adopting the recalibration of local clump absolute magnitudes in K , based on revised Hipparcos parallaxes ([Sect. 2.6](#)), the apparent magnitudes of the LMC clump stars yield for the true modulus $18.50 - 0.03 = 18.47 \pm 0.02$ (Laney et al. 2011) where the second term on the lhs is the population correction from (Salaris and Girardi 2002).

4.2.8 Tip of the Red Giant Branch

Sakai et al. (2000) obtained $I_o = 14.54 \pm 0.04$ and $(V - I)_o = 1.7$ for the LMC which with [Eq. 16.23](#) leads to a modulus of 18.68. It should be noted that the zero point of [Eq. 16.23](#) is not independently estimated but relies on other indicators (i.e., RR Lyrae stars)

The Cioni et al. (2000) measure of the TRGB in K_s together with their estimate of $[M/H]$ (-0.7), and [Eq. 16.24](#) gives a modulus of 18.50. Following the discussion in [Sect. 2.7](#), the standard error has been kept ~ 0.10 mag.

4.2.9 δ Sct Stars

McNamara et al. (2007) have used [Eq. 16.21](#) including the metallicity term and assuming $[Fe/H] = -0.4$ to obtain two estimates of the LMC modulus from δ Sct stars, 18.46 ± 0.19 and 18.50 ± 0.22 , and an estimated based on theoretical light curve shapes (18.48 ± 0.15).

4.2.10 Summary

The various moduli listed in [Table 16-5](#) scatter about a value close to 18.50 (depending on how they are weighted). The scatter can reasonably be accounted for by a combination of observational and absolute magnitude uncertainties together with possible uncertainties arising from the structure of the LMC in depth. It should be noted that the result for the red clump relies on theoretical population corrections and that the TRGB result may also be affected by population effects (see [Sect. 2.7](#)).

4.3 The Fornax Dwarf Spheroidal

The Fornax Dwarf Spheroidal is one of the nearer members of the Local Group and was the second one discovered by Shapley (1938) soon after his discovery of the Sculptor Dwarf Spheroidal. While there is a considerable range of properties among the dwarf spheroidals, this section should give some idea of the problems (and successes) in using various distance indicators in these systems which do not contain young stars (no classical Cepheids). Fornax contains both very old (~ 10 Gyr) and also intermediate age objects and a range of metallicities (e.g., Saviane et al. 2000; Tolstoy et al. 2003).

■ **Table 16-6**

Modulus of the Fornax Dwarf Spheroidal

Method	Modulus	Notes
TRGB(K)	20.63	[Fe/H] = -1.4
TRGB(K)	20.89	[Fe/H] = -1.0
TRGB(I)	20.91	
RR Lyraes	20.93	clusters
Clump(K)	20.71	corrected
Clump(I)	20.72	corrected
C-Miras	20.80	LMC = 18.50
δ Sct	20.69	[Fe/H] = -1.4
δ Sct	20.76	[Fe/H] = -1.0

► *Table 16-6* lists estimates of distance moduli. These have all been reduced to the scales discussed in earlier sections. A value of $E(B - V) = 0.03$ is assumed throughout. In view of uncertainties both in calibration and in metallicity and population corrections, the standard errors of the various moduli are difficult to assess but are probably ~ 0.1 .

4.3.1 TRGB(K)

As pointed out in ► *Sect. 2.7*, there are difference between observers on the apparent magnitude of the RGB tip at K_s . The results in ► *Table 16-6* adopt a mean of the values from Gullieuszik et al. (2007) and Pietrzyński et al. (2009a). There is also uncertainty in the correct metallicity to adopt. Gullieuszik et al. adopt $[Fe/H] = -1.4$, while Pietrzyński et al. use -1.0 . Adopting the absolute calibration of ► *Eq. 16.24*, one obtains the moduli listed for the two metallicities.

4.3.2 TRGB(I)

The tabulated result is derived from the V, I data of Rizzi et al. (2007b) together with ► *Eq. 16.23*.

4.3.3 RR Lyrae Variables

There have been a number of determinations of mean V for old HB stars in Fornax both for the field and for five globular clusters. There is a fair range in values ($\Delta V = 0.18$). Some of this may be due to a real depth effect. This would not be unexpected given the probable size and relative closeness of the system. A spread could also be caused by a spread in metallicities. The metallicity of the old field is uncertain, though it is generally assumed to be low ($[Fe/H] \sim -1.8$ to -2.0). It is perhaps safest to use the three clusters (numbers 1, 2, and 3) with spectroscopic abundances from Letarte et al. (2006) ($[Fe/H] -2.5, -2.1, -2.4$) and the mean V measurements of their RR Lyrae variables by Mackey and Gilmore (2003). With the calibration of ► *Sect. 2.3*, these give a mean modulus of 20.93.

4.3.4 Red Clump

The K_s apparent magnitude of the red clump is 19.21 (Pietrzyński et al. 2003, Gullieuszik et al. 2007) and 20.29 in I (Rizzi et al. 2007a). With the results in Sect. 2.6, these give moduli of 20.81 and 20.47 uncorrected for any population differences between Fornax and local Galactic field stars. Gullieuszik et al. (2007) calculate a population correction to the modulus of $\Delta K = -0.10$ based on the theoretical work of Salaris and Girardi (2002) together with an age–metallicity relation (Pont et al. 2004) and a star-formation history (Tolstoy et al. 2003). Similarly, Rizzi et al. (2007b) derive, $\Delta I = +0.25$.

4.3.5 Miras

Applying Eq. 16.20 (which depends on an LMC modulus of 18.50) to six C-Miras in Fornax (Whitelock et al. 2009) gives a modulus of 20.80 ± 0.09 .

4.3.6 δ Sct Stars

There is some uncertainty as to whether or not Eq. 16.21 applies to the Fornax δ Sct stars (Poretti et al. 2008). The results in Table 16-5 are based on comparing this equation with the Poretti et al. data and should be taken as approximate.

4.3.7 Summary

As Table 16-6 indicates a significant amount of scatter is introduced into the Fornax distance estimation by uncertainties in population corrections. These are likely to remain the main source of uncertainty for several indicators (clump, TRGB), especially when the population of a galaxy, such as a dwarf spheroidal, is only imperfectly known. A straight mean of the results in Table 16-6 give a modulus of 20.78.

5 Conclusions

The results discussed above suggest that the main stellar distance indicators currently in use are consistent at about the ~ 0.1 mag level or better. Provided care is taken, it is possible to derive distances within our own Galaxy and to nearby galaxies at this level of accuracy. However, population (e.g., age and metallicity) effects are still uncertain in several cases. Cepheids and RR Lyrae variables are currently the most precise indicators with type II Cepheids and Mira variables holding out considerable promise for the future. These may then be used to calibrate more general indicators (e.g., the Tully–Fisher relation, SNIa, etc.) which are beyond the scope of this chapter. A major step forward is likely in the near future from ground and space based astrometric projects. These will be the basis both for a significant improvements in the extragalactic distance scale and for a much more detailed study of the distribution of stars of all types in our Galaxy.

Appendix: Bias Correction

A.1 Introduction

There is a large literature on the statistical treatment of data and in particular on the recognition and correction for bias. This appendix is a summary to illustrate the main ways bias can occur in the establishment and use of distance scales and to indicate how one can correct, approximately, for such effects. Some of these results are used in the text.

A.2 Correction of Bias in Distance Moduli

The simplest case is for objects whose absolute magnitudes are distributed with a gaussian dispersion σ_{int} about a mean, M_m . Suppose these objects have a constant space density, interstellar absorption is negligible and the apparent magnitudes have negligible errors. If one selects objects with the same estimated distance modulus, some will have overestimated and some underestimated true moduli. However, since the number of stars per unit change in modulus increases with modulus, there will be more underestimated than overestimated moduli in this sample. Thus, the mean is too small and a correction for this bias is required. This case will be considered quantitatively below. However, a more general case is discussed first.

Consider a group of stars of a given type (say OB stars) distributed in space, for each of which one can estimate an individual absolute magnitude, M , from, say, spectroscopic luminosities. Assume for simplicity that each M is determined with the same gaussian observational uncertainty, σ_{obs} , due, for instance, to errors in measuring luminosity sensitive line ratios or assigning luminosity classes in the MK system. Assume also that the mean relation between M and the measured quantity has an intrinsic uncertainty in M of σ_{int} . If the apparent magnitude, m , of a star corrected for interstellar absorption is m_o (assumed to have a standard error σ_{m_o}), then the estimated distance modulus, $x = (m_o - M)$ differs from the true modulus, $x_c = x - \epsilon$, due to an error ϵ in x . With $(\sigma_{m_o}^2 + \sigma_{obs}^2 + \sigma_{int}^2) = \sigma_1^2$ and $h_1 = 1/\sigma_1\sqrt{2}$, the observed frequency distribution of x , for this group of stars, $P(x)$, is related to the true distribution, $Q(x - \epsilon)$ by,

$$P(x, \epsilon) = Q(x - \epsilon, \epsilon) = Q(x - \epsilon)(h_1/\sqrt{\pi})e^{-h_1^2\epsilon^2} \quad (\text{A.1})$$

and

$$P(x) = (h_1/\sqrt{\pi}) \int_{-\infty}^{+\infty} Q(x - \epsilon)e^{-h_1^2\epsilon^2} d\epsilon \quad (\text{A.2})$$

The mean value of ϵ , $\bar{\epsilon}$, is given by,

$$\bar{\epsilon} \int_{-\infty}^{+\infty} Q(x - \epsilon)e^{-h_1^2\epsilon^2} d\epsilon = \int_{-\infty}^{+\infty} \epsilon Q(x - \epsilon)e^{-h_1^2\epsilon^2} d\epsilon \quad (\text{A.3})$$

When integrated by parts the rhs of this equation becomes,

$$-\frac{1}{2h_1^2} \left[Q(x - \epsilon)e^{-h_1^2\epsilon^2} \right]_{-\infty}^{+\infty} + \frac{1}{2h_1^2} \int_{-\infty}^{+\infty} e^{-h_1^2\epsilon^2} \frac{d}{d\epsilon} Q(x - \epsilon) \cdot d\epsilon$$

The integrated part vanishes at both limits. In addition,

$$\frac{d}{d\epsilon} \cdot Q(x - \epsilon) = -\frac{d}{dx} \cdot Q(x - \epsilon). \quad (\text{A.4})$$

Thus, the remaining term becomes

$$-\frac{1}{2h_1^2} \frac{d}{dx} \int_{-\infty}^{+\infty} Q(x - \epsilon) e^{-h_1^2 \epsilon^2} d\epsilon$$

Hence, using A2,

$$\bar{\epsilon} = -\sigma_1^2 P'(x)/P(x) \quad (\text{A.5})$$

and

$$\bar{x}_c = x + \sigma_1^2 P'(x)/P(x) \quad (\text{A.6})$$

where here and elsewhere primes denote differentials and bars indicate means or statistically corrected values.

In the above, the absolute magnitudes of the objects under study were estimated individually. An important special case arises when the objects are assumed to belong to a class of mean absolute magnitude, M_m and dispersion σ_{int} . An argument along the lines given above, but with $\sigma_{obs} = 0$, leads to a relation similar to [Eq. A.5](#). In that relation, $P(x)$ is the apparent frequency distribution of the objects actually included in the study and, in this case, $P(x) \equiv P(m_o)$. This will generally be a subset of all the objects of this type whose apparent frequency distribution is $\nu(x)$. Provided the interstellar absorption is negligible and the selection of objects from $\nu(x) \equiv \nu(m)$ is done solely on the basis of apparent magnitude, then

$$P(x) \equiv P(m) = f(m)\nu(m). \quad (\text{A.7})$$

Here, $f(m)$ is the fraction of $\nu(m)$ actually in the selected sample. Similar results apply if the interstellar absorption is constant. Then,

$$P'(m)/P(m) = \nu'(m)/\nu(m) + f'(m)/f(m). \quad (\text{A.8})$$

Evidently, choosing stars by apparent magnitude implies that $f(m)$ is a maximum at the chosen magnitude. Thus, the final term on the rhs of [Eq. A.8](#) is zero, and the correction depends directly on the underlying apparent space distribution and becomes

$$\bar{\epsilon} = -\sigma_{int}^2 \nu'(m)/\nu(m) \quad (\text{A.9})$$

This relation can also be obtained directly by noting that in this case, one can make the substitutions, $P(x, \epsilon) = f(m)\nu(x, \epsilon)$ and $Q(x - \epsilon, \epsilon) = f(m)u(x - \epsilon, \epsilon)$ in [Eq. A.1](#). Here, $u(x)$ is the true underlying frequency distribution. The difference between [Eq. A.9](#) which applies to objects of the same mean absolute magnitude and [Eq. A.5](#) has to be kept in mind. Using [Eq. A.9](#) when in fact [Eq. A.5](#) is required will lead to incorrect bias corrections.

If it is assumed that the objects to which [Eq. A.9](#) applies are distributed in space according to some power law in their true distances (r_c), r_c^n ($n = 2$ for a constant density distribution), then the true frequency distribution is $u(x) \propto e^{bx(n+1)}$ where $b = 0.46$. The relation between $\nu(x)$ and $u(x - \epsilon)$ is given by an equation of the same form as [Eq. A.2](#). For a power-law density distribution, this can be solved explicitly and shows that $\nu(x)$ is also proportional to $e^{bx(n+1)}$ and hence,

$$\bar{\epsilon} = -b(n+1)\sigma_{int}^2 \quad (\text{A.10})$$

and in the case of a constant space density,

$$\bar{\epsilon} = -1.38\sigma_{int}^2 \quad (\text{A.11})$$

This is the well-known relation given by Eddington (1914) and also by Malmquist (1920).

Equation 16.9–16.11 are, under the various restrictions listed above, also appropriate when M is estimated from an auxiliary quantity, provided the error arising from uncertainty in this quantity, σ_{obs} , is negligible. This follows since if the auxiliary quantity is measured with negligible error, the sample of objects can be notionally divided into groups at each value of this quantity each with its own mean M and σ_{int} . Estimation of the luminosities of variable stars from their periods can be an example of this since the periods are usually well determined. If σ_{obs} is significant, this cannot be done (see Sect. A.3 for further discussion).

A related problem is for objects of the class just discussed which are all at the same true distance, in, say, a cluster or external galaxy. In that case, the distribution of the apparent moduli, $v(x) \equiv v(m)$, about the true modulus would be gaussian. Hence, the mean value of the first term on the rhs in Eq. A.8 for such a sample would be zero. Thus, the mean value of $\bar{\epsilon}$, ($\bar{\bar{\epsilon}}$), for the actual sample of objects observed depends entirely on $f(m)$. This situation is sometimes considered separately from the case of objects distributed in space. Teerikorpi (1997) refers to these two cases as Malmquist bias of the first and second kind. However, they are both applications of Eqs. A.5 and A.8.

In practice, other selection effects may be present. For instance, stars may be selected by proper motion for parallax measurement (see, e.g., Hanson 1979), and differential interstellar extinction complicates the effects of magnitude selection. In the Tully–Fisher relation, there may be biases due to diameter selection effects. The methods employed to deal with these problems in particular cases will depend on the size of the observational and intrinsic dispersions involved and also on estimated effects of magnitude selection. Some methods applicable to the case when $f(m)$ is constant up to a certain value of m and zero for larger values (i.e., magnitude truncation) are discussed in Teerikorpi (1997). These include the “triple-entry” correction method (Sandage 1994a, b; Spaenhauer 1978).

A.3 Bias in Distances from Absolute Magnitudes

It frequently occurs in Galactic and extragalactic work that one has individually measured absolute magnitudes or an adopted absolute magnitude for a class of objects. What may then be required are bias corrections applicable to some power (s) of the estimated distance (r) rather than to the distance modulus itself. In that case, if r_c is the true distance of an object, then $r_c^s = r^s e^{-bes}$ where ϵ is as before the error in the modulus. Evidently, $\overline{e^{-bes}}$ is required, and this is found by replacing ϵ and $\bar{\epsilon}$ by e^{-bes} and $\overline{e^{-bes}}$ in Eq. A.3. The integral is solved by making the substitution $\eta = \epsilon + bs\sigma^2$, where σ is the relevant value of the dispersion for the various cases discussed in Sect. A.2. It is found that, (see Feast 1972).

$$\overline{r_c^s} = r^s e^{0.5b^2s^2\sigma^2} P(x + bs\sigma^2)/P(x) \quad (\text{A.12})$$

The importance of corrections of this kind and the need to determine $P(x)$ in any sample under study was demonstrated by Feast and Shuttleworth (1965) in their study of the Galactic kinematics of OB-type stars from radial velocities. For such stars, there is a wide range in luminosities which are generally estimated from MK spectral types and luminosity classes or from the strengths of Balmer lines. High-luminosity OB stars are much less frequent in space than lower luminosity ones. Thus, at large distances, in magnitude limited samples, there will be more overestimated distances than underestimated ones. Using a series expansion of Eq. A.12, it

was estimated that in the sample of stars under study, this overestimation amounted to about 2 kpc at an uncorrected distance of 8 kpc.

If the objects in question have a gaussian distribution about some mean absolute magnitude, it follows, as in [Sect. A.2](#), that the correction depends only on the underlying apparent space density and thus on $v(x + bs\sigma_{int}^2)/v(x)$. For a power-law space distribution as in [Sect. A.2](#), the correction is then

$$\bar{r}_c = r^s e^{b^2\sigma_{int}^2(0.5s^2+s(n+1))}. \quad (\text{A.13})$$

In any application, it is important to take the correct value of s . For instance, consider the case of a complete survey down to some apparent magnitude for a system of constant space density (N). Then since the number of stars observed per unit observed distance is, $Nr^2 dr$, application of [Eq. A.13](#) shows that the N will be overestimated by $e^{2.22\sigma^2}$ or, if the intrinsic dispersion, σ , is 0.5, by a factor of 1.74. The apparent value of N may, of course, be derived from the observed distribution of distance moduli and the correction then applied.

Note that both the Oort constant of differential galactic rotation, A , and the Hubble constant, H_0 , can, to a first approximation, be derived from a radial velocity divided by a distance so that $s = -1$ is needed in these cases.

A.4 Bias in Mean Parallaxes and Absolute Magnitudes

The discussion in [Sects. A.2](#) and [A.3](#) assumes that the uncertainty in the absolute magnitude (and distance modulus), either measured or assumed, is gaussian. This will not in general be the case when M or x is derived as a function of some measured quantity. The derivation of distance moduli from parallaxes or from the Tully–Fisher relation via line-widths are examples of this. Here, the case of parallaxes will be discussed, but similar results obviously apply in other analogous cases.

Consider a group of stars with a mean absolute magnitude, M_m , distributed in space, and for each of which there is a measured parallax ω . Assume also that the standard error of ω , σ_ω , is the same for all the stars. The correction for the expected mean bias in ω can obviously be obtained from the results in [Sect. A.2](#) with x replaced by ω in the equations and the equivalent of [Eq. A.5](#) applies. Note in particular that $\bar{\epsilon}$ may be either positive or negative depending on $P(\omega)$. Evidently, a distance modulus and, hence, an absolute magnitude can be derived from the bias corrected parallax. Because this involves the conversion from natural to logarithmic quantities, this requires a (non-gaussian) correction which to a first approximation is $+1.085(\sigma_\omega/\omega)^2$ in the distance modulus. These corrections lead to an estimate of the true distance modulus.

If the stars are selected entirely on the basis of their parallaxes from the underlying distribution, then a form of [Eq. A.9](#) applies with m replaced by ω and the appropriate σ .

In a frequently quoted paper, Lutz and Kelker (1973) approach the problem just discussed by assuming that the stars are selected only by parallax and that the true density distribution of the class of objects concerned is constant. That is, $u(\omega) \propto \omega^{-4}$. They then compute (numerically) the correction to a derived absolute magnitude as a function of σ_ω/ω . They found that in this case, their method broke down for $\sigma_\omega/\omega > \sim 0.18$.

To compare the two approaches, it is necessary to derive $v'(\omega)/v(\omega)$ from the $u(\omega)$ adopted by Lutz and Keller. This can be done by expanding $(\omega - \epsilon)^{-4}$ as a Taylor series in the equivalent of [Eq. A.2](#) and taking the first few terms for $v(\omega)$. One then obtains, after applying the factor

for conversion from natural to logarithmic quantities, numerical results from the first method which are close to those of Lutz and Kelker. There is a problem of convergence of the Taylor series when an infinite uniform density is assumed for the true distribution, and this may be the reason that the Lutz and Kelker method fails above a certain value of σ_ω/ω .

In general, the bias corrections are non-gaussian and asymmetrical. One might expect that this problem would be particularly significant, where true distributions are supposed to extend indefinitely outward from the Sun. Koen (1992) showed that for the Lutz–Kelker case as well as for the case $u(\omega) \propto \omega^{-2}$, the distributions of errors in the absolute magnitude corrections become highly skew at quite moderate value of σ_ω/ω . For instance, in the Lutz–Kelker case for $\sigma_\omega/\omega = 0.15$, the mean correction to the absolute magnitude is -0.28 mag, the most likely correction (peak of the distribution) is -0.16 mag, whilst the upper and lower 90% confidence limits are -1.00 mag and $+0.32$ mag.

The methods discussed in this subsection are important because they can be, and in the case of the Lutz–Kelker formulation often have been, applied to single objects. There are however two caveats that should be mentioned. First, one often wants to obtain a mean absolute magnitude by combining the results from several objects and the proper weighting of objects with different and asymmetrical uncertainties becomes problematic. Secondly, there is the problem of knowing the underlying distribution from which the parallaxes are drawn. The assumption that the selection is made entirely on the basis of the measured quantity (parallax), as in the example just discussed, is unlikely to be generally true. For instance, in most cases, magnitude selection must affect the result. If this could be quantified, it could be built into a frequentist argument or a Bayesian prior. Generally, this cannot be done. However, in most cases, neglecting magnitude selection will result in overcorrection of the absolute magnitudes. Serious problems then only arise if the corrections are large.

A.5 Determination of Absolute Magnitudes: The Reduced Parallax Method

Consider the case of the measured parallaxes of objects with a mean absolute magnitude M_o per unit volume which one wishes to estimate. Assume that the intrinsic scatter in M_o , σ_{int} is known (perhaps from a set of such objects in a cluster or galaxy). Provided the objects have not been selected on the basis of their measured parallaxes (ω) or ω/σ_ω , some of the problems mentioned in the last subsection can be avoided by combining the results in parallax space. This is the method of reduced parallaxes which scales the measured parallaxes to the values they would have at the same dereddened apparent magnitude.

The version given here follows Feast (2002). A detailed study with a nonlinear version of the method is given by Smith (2003). In any application, it is important to judge the method most appropriate for the material at hand and the known properties of the objects under study.

If the reddening corrected magnitude is m_o , an estimate of the quantity, $10^{0.2M}$ is given by

$$\overline{10^{0.2M}} = \sum 0.01\omega 10^{0.2m_o} p / \sum p \quad (\text{A.14})$$

Where the parallax is in milliarcsecs and p is the weight of a star given by

$$(0.01\sigma_T 10^{0.2m_o})^2 = 1/p \quad (\text{A.15})$$

and where

$$\sigma_T^2 = \sigma_\omega^2 + b^2 \omega_{M_o} \sigma_2^2 \quad (\text{A.16})$$

and

$$\sigma_2^2 = \sigma_{int}^2 + \sigma_{m_o} \quad (\text{A.17})$$

Here, ω_{M_o} is the photometric parallax of the object obtained from $(m_o - M_o)$, where M_o is obtained, if necessary, by iteration. **Equation A.16** is discussed by Koen and Laney (1998). **Equation A.14** only gives an unbiased estimate of $10^{0.2M_o}$ if σ_2 is negligible. In general, if a photometric distance modulus has an error ϵ , then

$$\overline{10^{0.2M_o}} = \overline{10^{0.2(M-\epsilon)}} = \overline{e^{bM}} \cdot \overline{e^{-b\epsilon}} \quad (\text{A.18})$$

and following methods similar to those outlined in earlier subsections one finds after correction from exponentials to natural numbers and for Eddington–Malmquist bias on the assumption of a constant space density, that,

$$M_o = 5 \log(\overline{10^{0.2M}}) + 1.151\sigma_2^2 - 0.23\sigma_o \quad (\text{A.19})$$

where σ_o is the observational error in the derived value of M_o . This equation demonstrates the usefulness of the reduced parallax method. It may usually be assumed that the error is small in the reddening corrected magnitude (or that a reddening free index has been used), and for good distance indicators, the intrinsic scatter in absolute magnitude will be small. In that case, the second term on the rhs is small and, in some cases, negligible. Also, for practical distributions, the coefficient of this term is likely, as has already been seen, to be less than 1.151. If this term is negligible, then the use of the final term can be avoided. $\overline{10^{0.2M}}$ is then the mean-derived parallax for a standard apparent magnitude, and this can be used directly with photometric parallaxes which define the relative scale to obtain what are effectively parallaxes of distant stars of the type under consideration.

The coefficient of σ_2^2 in **Eq. A.19** assumes a uniform space density. The appropriate value for other assumptions can be obtained from, for example, **Eq. A.13**.

The following is a detailed example of the effects of adopting incorrect weighting systems. The work of Wegner (2007) in calibrating the MK system using Hipparcos parallaxes was mentioned in **Sect. 2.10.1**. His tabulated results depend on a method which is superficially similar to the method of reduced parallaxes. However, the weights, p , which he adopts are given by,

$$\sigma_\omega^2 = 1/p. \quad (\text{A.20})$$

Comparison with **Eq. A.15** shows that this gives higher weight to the fainter stars than the method of reduced parallaxes. This difference in weighting has a significant effect on the final results. For instance, with weighting as in A20 Wegner obtains $M_V = 5.29 \pm 0.07$ for KOV stars (where the standard error of the mean has been derived from his listed dispersion). Wegner also mentions that an estimate of $10^{0.2M_V}$ can be derived from the slope of the relation between ω and $10^{-0.2(V_o+5)}$. A little consideration shows that this is closer to the method of reduced parallaxes. An approximate measurement of the slope of this relation for KOV stars given in Figure 1 of Wegner (2007) yields $M_V = 5.88$ with an uncertainty of ~ 0.1 mag. That the statistical method adopted by Wegner leads to too bright a luminosity in this case is also shown by the four KOV MK system standard stars in the list of Garcia (1989) which have Hipparcos parallaxes. These are all of high quality (σ_ω/ω between 0.003 and 0.040) and yield $M_V = 5.65 \pm 0.08$. Note that none of the above estimates take into account the final two terms in **Eq. A.19**. The last term is not significant in this case, but the other one will result in a somewhat fainter M_V .

A.6 General Comments

The alternative to the approach taken above is to employ some form of simulating observations, i.e., modelling. One form of modelling is to use an inverse relation. This method was first proposed by Schechter (1980) for the Tully–Fisher relation and is discussed by Teerikorpi (1997). Monte Carlo simulations can also be used. Useful references on some of the problems discussed in this appendix are Pont (1999), Smith (2003, 2006); and papers to which these authors refer.

Acknowledgements

Dr. Noriyuki Matsunaga very kindly prepared [▶ Fig. 16-3](#) and also read and commented on a draft version of the chapter. I am grateful to Prof. D. Lynden-Bell for comments on the Appendix.

Cross-References

- ▶ [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- ▶ [Dynamics of Disks and Warps](#)
- ▶ [History of Dark Matter in Galaxies](#)
- ▶ [Mass Distribution and Rotation Curve in the Galaxy](#)

References

- Alloin, D., & Gieren, W., eds. 2003, *Stellar Candles for the Extragalactic Distance Scale* (Berlin: Springer)
- An, D., Terdrup, D. M., & Pinsonneault, M. H. 2007, The distances of open clusters from main sequence fitting. IV. Galactic Cepheids, the LMC, and the local distance scale, *ApJ*, 671, 1640–1668
- Arp, H. C. 1956, Novae in the Andromeda nebula, *AJ*, 61, 15–34
- Balona, L., & Crampton, D. 1974, The $H\gamma$ -absolute magnitude calibration, *MN*, 166, 203–217
- Bellazzini, M. 2008, The tip of the red giant branch, *Mem Soc Ast It*, 79, 440–447
- Benedict, G. F., & McArthur, B. E. 2004, High-precision stellar parallaxes from the Hubble Space Telescope fine guidance sensors, in *Transits of Venus; New Views of the Solar System and the Galaxy*, eds. D. W. Kurtz et al. (Cambridge: Cambridge University Press), 333–346
- Benedict, G. F., et al. 2007, Hubble space telescope fine guidance sensor parallaxes of galactic cepheid variables, *AJ*, 133, 1810–1827
- Benedict, G. F., et al. 2011, Calibrating the distance scale with Galactic RR Lyrae star parallaxes, *AJ*, 142:187
- Bird, J. C., Stanek, K. Z., & Prieto, J. L. 2009, Using ultra long period Cepheids to extend the cosmic distance ladder to 10 Mpc and beyond, *ApJ*, 695, 874–882
- Blaauw, A. 1963, The calibration of luminosity criteria, in *Basic Astronomical Data*, ed. K. Aa. Strand (Chicago: University of Chicago Press), 383–420
- Blommaert, J. A. D. L., et al. 1998, The nature of OH/IR stars in the Galactic Centre, *A&A*, 329, 991–1009
- Borissova, J., et al. 2009, Properties of RR lyrae stars in the inner regions of the Large Magellanic Cloud. *A&A*, 502, 505–514
- Bresolin, F. 2003, Blue supergiants as tools for extragalactic distances – empirical diagnostics, in *Stellar Candles for the Extragalactic Distance Scale*, eds. D. Alloin, W. Gieren (Berlin: Springer), 149–174
- Brown, A. G. A., et al. 2000, OB associations the Hipparcos view, in *Star Formation from the Small to*

- the Large Scale, eds. Favata et al. (ESA SP-445, Paris: European Space Agency), 239–248
- Cacciari, C., & Clementini, G. 2003, Globular cluster distances from RR Lyrae stars, in *Stellar Candles for the Extragalactic Distance Scale*, eds. D. Alloin, W. Gieren (Berlin: Springer), 105–122
- Carpenter, J. M. 2001, Color transformations for the 2MASS second incremental data release, *AJ*, 121, 2851–2871, and later versions in <http://www.ipac.caltech.edu/2mass/>
- Carretta, E., et al. 2000, Distances, ages and epoch of formation of globular clusters, *ApJ*, 533, 215–235
- Cioni, M. R., et al. 2000, The tip of the red giant branch distance of the Magellanic Clouds: Results from the DENIS survey, *A&A*, 359, 601–614
- Clausen, J. V. 2004, Eclipsing binaries as precise standard candles, *New Ast Rev*, 48, 679–685
- Clausen, J. V., et al. 2003, Eclipsing binaries in the Magellanic Clouds. *uvby* CCD light curves and photometric analyses for HV982(LMC), HV12578(LMC), HV1433(SMC) and HV11284(SMC), *A&A*, 402, 509–530
- Clementini, G., et al. 2005, The metal abundance distribution of the oldest stellar component in the Sculptor dwarf spheroidal galaxy, *MN*, 363, 734–748
- Crawford, D. L. 1978, Empirical calibrations of the *uvby* systems. II. The B-type stars, *AJ*, 83, 48–63
- Dall’Ora, M., et al. 2004, The distance of the LMC cluster Reticulum from the *K*-band period-luminosity-metallicity relation of RR Lyrae stars, *ApJ*, 610, 269–274
- Dambis, A. K. 2009, The kinematics and zero-point of the $\log P - \langle M_K \rangle$ relation for Galactic RR Lyrae variables via statistical parallax, *MN*, 396, 553–569
- Degl’Innocenti, S., et al. 2004, The $^{14}N(p, \gamma)^{15}O$ reaction, solar neutrinos and the age of globular clusters, *PhLB*, 590, 13–20
- Dehnen, W., & Binney, J. J. 1998, Local stellar kinematics from Hipparcos data, *MN*, 298, 387–394
- Delhaye, J. 1965, Solar motion and velocity distribution of common stars, in *Galactic Structure*, eds. A. Blaauw, M. Schmidt (Chicago: University of Chicago Press), 61–84
- de Zeeuw, P. T., et al. 1999, A Hipparcos census of the nearby OB associations, *AJ*, 117, 354–399
- Di Benedetto, G. P. 2008, The Cepheid distance to the Large Magellanic Cloud and NGC4258 by surface brightness technique and improved calibration of the cosmic distance scale, *MN*, 390, 1762–1776
- Downes, R. A., & Duerbeck, H. W. 2000, Optical imaging of nova shells and the maximum magnitude-rate of decline relationship, *AJ*, 120, 2007–2037
- Eddington, A. S. 1914, *Stellar Movements and the Structure of the Universe* (London: Macmillan), 172
- Eyre, A., & Binney, J. 2009, Fitting orbits to tidal streams with proper motions, *MN*, 399, L160–L163
- Feast, M. W., & Thackeray, A. D. 1958, Analysis of radial velocities of distant B-type stars, *MN*, 118, 125–153
- Feast, M. W., & Shuttlesworth, M. 1965, The kinematics of B stars, Cepheids, Galactic Clusters and interstellar gas in the Galaxy, *MN*, 130, 245–280
- Feast, M. W. 1972, A problem in distance determination for Mira variables with an Appendix on OB-star distances, *Vistas Astron*, 13, 207–221
- Feast, M. W., & Walker, A. R. 1987, Cepheids as distance indicators, *ARAA*, 25, 345–375
- Feast, M. W., & Catchpole, R. M. 1997, The Cepheid period-luminosity zero-point from Hipparcos trigonometrical parallaxes, *MN*, 286, L1–L5
- Feast, M. W., & Whitelock, P. A. 1997, Galactic kinematics of Cepheids from Hipparcos proper motions, *MNRAS*, 291, 683–693
- Feast, M. W. 1999a, Pulsating stars in globular clusters and their use, in *Globular Clusters*, eds. C. Roger Martínez et al. (Cambridge: Cambridge University press), 251–290
- Feast, M. W. 1999b, Cepheids as distance indicators, *PASP*, 111, 775–793
- Feast, M. W. 2000, The local solar motion and the scale length of the Galactic disc, *MN*, 313, 596–598
- Feast, M. W. 2002, Bias in absolute magnitude determination from parallaxes, *MN*, 337, 1035–1037
- Feast, M. W. 2003, Current uncertainties in the use of Cepheids as distance indicators, in *Stellar Candles for the Extragalactic Distance Scale*, eds. D. Alloin, W. Gieren (Berlin: Springer), 45–70
- Feast, M. W., Whitelock, P. A. & Menzies, J. W. 2006, Carbon-rich Mira variables: kinematics and absolute magnitudes, *MN*, 369, 791–797
- Feast, M. W. et al. 2008, The luminosities and distance scales of type II Cepheid and RR Lyrae variables, *MN*, 386, 2115–2134
- Feast, M. W. 2009, The ages, masses, evolution and kinematics of Mira variables, in *Nat. Obs. Jap. Tokyo. AGB variables and related phenomena Tokyo*, eds. T. Ueta et al. (ArXiv:0812.0250), 48–52
- Fernie, J. D. 1969, The period-luminosity relation: A historical review, *PASP*, 81, 707–731

- Fernley, J., et al. 1998, The absolute magnitudes of RR Lyraes from Hipparcos parallaxes and proper motions, *A&A*, 330, 515–520
- Ferraro, F., et al. 2000, A new infrared array survey of galactic globular clusters *AJ*, 119, 1282–1295
- FitzPatrick, E. L., et al. 2003, Fundamental properties and distances of the Large Magellanic Cloud eclipsing binaries. IV. HV5936, *ApJ*, 587, 685–700
- Fouqué, P., et al. 2007, A new calibration of Galactic Cepheid period-luminosity relations from B to K bands, and a comparison to LMC relations, *A&A*, 476, 73–81
- Fusi Pecci, F., et al. 1996, The M_V^{HB} versus [Fe/H] calibration. I. *HST* colour-magnitude diagrams of eight globular clusters in M31, *AJ*, 112, 1461–1471
- Garcia, B. 1989, A list of MK standard stars. *Bull inf, CDS*, 36, 27–90
- Ghez, A. M., et al. 2008, Measuring distance and properties of the Milky Way's central super-massive black hole with stellar orbits, *ApJ*, 689, 1044–1062
- Gillessen, S., et al. 2009a, Monitoring stellar orbits around the massive black hole in the Galactic Centre, *ApJ*, 692, 1075–1109
- Gillessen, S., et al. 2009b, The orbit of the star S2 around SgrA* from the VLT and Keck data, *ApJ*, 707, L114–L117
- Gilmozzi, R., & Della Valle, M. 2003, Novae as distance indicators, in *Stellar Candles for the Extragalactic Distance Scale*, eds. D. Alloin, W. Gieren (Berlin: Springer), 229–242
- Gingold, R. 1985, The evolutionary status of type II Cepheids, *Mem Soc Ast It*, 56, 169–192
- Girardi, L. 1999, A secondary clump of red giant stars: Why and where, *MN*, 308, 818–832
- Girardi, L., & Salaris, M. 2001, Population effects on the red giant clump absolute magnitude, and distance determination to nearby galaxies, *MN*, 323, 109–129
- Girardi, L., Rubele, S., & Kerber, L. 2009, Discovery of two distinct red clumps in NGC419: a rare snapshot of a cluster at the onset of degeneracy, *MN*, 394, L74–L78
- Glass, I. S., & Feast, M. W. 1982, Infrared photometry of Mira variables in the Baade windows and the distance to the Galactic Centre, *MN*, 198, 199–214
- Glass, I. S., et al. 1995, Long-period variables in the SgrI field of the Galactic Bulge, *MN*, 273, 383–400
- Glass, I. S., & Lloyd Evans, T. 2003, The calibrating stars of the Mira P-L relation, *MN*, 343, 67–74
- Gratton, R. G., et al. 2003, Distances and ages of NGC6397, NGC6752 and 47 Tuc, *A&A*, 408, 529–543
- Gratton, R. G., et al. 2004, Metal abundances of RR Lyrae stars in the bar of the Large Magellanic Cloud, *A&A*, 421, 937–952
- Grocholski, A. J., et al. 2007, Distances of populous clusters in the Large Magellanic Cloud via K-band luminosity of the red clump, *AJ*, 134, 680–693
- Groenewegen, M. A. T., Udalski, A., & Bono, G. 2008, The distance of the Galactic Centre based on population II Cepheids and RR Lyrae stars, *A&A*, 481, 441–448
- Groenewegen, M. A. T. 2008, The red clump absolute magnitude based on revised Hipparcos parallaxes, *A&A*, 488, 935–941
- Gullieuszik, M., et al. 2007, Near-infrared observations of the Fornax dwarf galaxy. I. The red giant branch, *A&A*, 467, 1025–1036
- Hanson, R. B. 1979, A practical method to improve luminosity calibrations from trigonometric parallaxes, *MN*, 186, 875–896
- Hendry, M. A., O'dell, M. A., & Collier-Cameron, A. 1993, A new method for estimating the distance of young open clusters, *MN*, 265, 983–995
- Herrnstein, J. R., et al. 1999, A geometric distance to the galaxy NGC4258 from orbital motions in a nuclear gas disc, *Nat*, 400, 539–541
- Holmberg, J., Nordström, B., & Andersen, J., 2007, The Geneva-Copenhagen survey of the solar neighbourhood II, *A&A*, 475, 519–537
- Imbriani, G., et al. 2004, The bottleneck of CNO burning and the age of the globular clusters, *A&A*, 420, 625–629
- Ita, Y., et al. 2004, Variable stars in the Magellanic Clouds; Results from OGLE and Sirius, *MN*, 347, 720–728
- Ivezic, Z., et al. 2008, LSST: From science drivers to reference design and anticipated products, *astro-ph/0805.2366*
- Jurcevic, J. S., Pierce, M. J., & Jacoby, G. H. 2000, Period-luminosity relations for red supergiant variable-II. The distance of M101, *MN*, 313, 868–880
- Juric, M., et al. 2008, The Milky Way Tomography with SDSS. I. Stellar number density distribution, *ApJ*, 673, 864–914
- Karovska, M., et al. 1991, Asymmetries in the atmosphere of Mira, *ApJ*, 374, L51–L54
- Kato, D., et al. 2007, The IRSF Magellanic Cloud point source catalogue, *PASJ*, 59, 615–641
- Kerr, F. J., & Lynden Bell, D. 1986, Review of Galactic constants, *MN*, 221, 1023–1038

- Kinman, T. D., Mould, J. R., & Wood, P. R. 1987, Variable stars in local group galaxies. I-M33, *AJ*, 93, 833–850
- Klement, R. et al. 2009, Halo streams in the seventh Sloan digital sky survey data release, *ApJ*, 698, 865–894
- Koen, C. 1992, Confidence intervals for the Lutz-Kelker correction, *MN*, 256, 65–68
- Koen, C., & Laney, D. 1998, On the determination of absolute magnitude zero-points from Hipparcos parallaxes, *MN*, 301, 582–584
- Kudritzki, R.-P., & Przybilla, N. 2003, Blue supergiants as tools for extragalactic distances – theoretical concepts, in *Stellar Candles for the Extragalactic Distance Scale*, eds. D. Alloin, W. Gieren (Berlin: Springer), 123–149
- Lacour, S., et al. 2009, The pulsation of χ Cygni imaged by optical interferometry: A novel technique to derive distance and mass of Mira stars, arXiv:0910.3869
- Laney, C. D., & Stobie, R. S. 1995, The radii of Galactic Cepheids, *MN*, 274, 337–360
- Laney, C. D., Jonek, M. D., & Pietrzynski, G. 2011, A new LMC K band distance from precision measurements of nearby red clump stars. *MN* in press (ArXiv:1109.4800)
- Lane, B. F., et al. 2002, Long-baseline interferometric observations of Cepheids, *ApJ*, 573, 330–337
- Lattanzi, M. G., et al. 1997, Interferometric angular diameters of Mira variables with the Hubble Space Telescope, *ApJ*, 485, 328–332
- Layden, A. C. 1999, Absolute magnitudes derived using the statistical parallax method. In *Post-Hipparcos Cosmic Candles*, in eds. A. Heck, & F. Caputo (Dordrecht: Kluwer), 37–52
- Leavitt, H. A., & Pickering, E. C. 1912, Periods of 25 variables in the Small Magellanic Cloud, *Harv Obs Circ*, 173, p1–p3
- Letarte, B., et al. 2006, VLT/UVES spectroscopy of individual stars in three globular clusters in the Fornax dwarf spheroidal galaxy, *A&A*, 453, 547–554
- Lindgren, L., et al. 2008, The Gaia mission: science, organization and present status, in *IAU Symp.* 248, 217–223
- Longmore, A. J., Fernley, J. A., & Jameson, R. F. 1986, RR Lyrae stars in globular clusters—Better distances from infrared measurements? *MN*, 220, 279–287
- Lutz, T. E., & Kelker, D. H. 1973, On the use of trigonometrical parallaxes for the calibration of luminosity systems: Theory, *PASP*, 85, 573–578
- Mackey, A. D., & Gilmore, G. 2003, RR Lyrae stars in four globular clusters in the Fornax dwarf galaxy, *MN*, 345, 747–761
- Macri, L. M., et al. 2006, A new Cepheid distance to the maser-host galaxy NGC 4258 and its implications for the Hubble constant, *ApJ*, 652, 1133–1149
- Madore, B. F., & Freedman, W. L. 1995, The tip of the red giant branch as a distance indicator for resolved galaxies. 2. Computer simulations, *AJ*, 109, 1645–1652
- Makarov, D., et al. 2006, Tip of the red giant branch distances. I. Optimization of a maximum likelihood algorithm, *AJ*, 132, 2729–2742
- Malmquist, K. G. 1920, A study of the stars of spectral type A, *Lund Medd Ser II*, 22, 1–68
- Matsunaga, N., et al. 2006, The period-luminosity relation for type II Cepheids in globular clusters, *MN*, 370, 1979–1990
- Matsunaga, N., et al. 2009a, Period-luminosity relations for type II Cepheids and their application, *MN*, 397, 933–942
- Matsunaga, N., et al. 2009b, A near-infrared survey of Miras and the distance to the Galactic Centre, *MN*, 399, 1709–1729
- Matsunaga, N., et al. 2011, Three classical Cepheid variable stars in the nuclear bulge of the Milky Way, *Nature*, 477, 188–190
- McMillan, P. J., & Binney, J. J. 2010, The uncertainties in Galactic parameters, *MNRAS*, 402, 934
- McNamara, D. H. 1999, The slope of the RR Lyrae variables $M_v f([Fe/H])$ relation, *PASP*, 111, 489–493
- McNamara, D. H., Clementini, G., & Marconi, M. 2007, A δ Sct distance to the Large Magellanic Cloud, *AJ*, 133, 2752–2763
- Mitchell, R. C., et al. 2002, Detailed spectroscopic analysis of SN1987A: The distance to the Large Magellanic Cloud using the spectral-fitting expanding atmosphere method, *ApJ*, 574, 293–305
- Mihalas, D. 1968, *Galactic Astronomy* (Chap. 6; San Francisco: Freeman and Co.)
- Miyamoto, M., & Zhu, Z. 1998, Galactic interior motions derived from *Hipparcos* proper motions. I. young disk population, *AJ*, 115, 1483–1491
- Morgan, W. W., Keenan, P. C., & Kellman, E. 1943, *An Atlas of Stellar Spectra* (Chicago: University of Chicago Press)
- Morgan, W. W., Whitford, A. E., & Code, A. D. 1953, *Studies in Galactic structure. I. A preliminary determination of the space distribution of blue supergiants*, *ApJ*, 118, 318–322
- Morgan, W. W., & Keenan, P. C. 1973, Spectral classification, *ARAA*, 11, 29–50
- Murray, C. A. 1983, *Vectorial Astrometry* (Bristol: Hilger)

- Nakagawa, A., et al. 2009, Period-luminosity relation of the Galactic Mira variables measured with Vera, in eds. T. Ueta et al. Tokyo. AGB stars and related phenomena, 58–61
- Ngeow, C.-C., & Kanbur, S. M. 2005, The linearity of the Wesenheit function for the Large Magellanic Cloud Cepheids, *MN*, 360, 1033–1039
- Ngeow, C.-C., et al. 2005, Further empirical evidence for the non-linearity of the period-luminosity relations as seen in the Large Magellanic Cloud Cepheids, *MN*, 363, 831–846
- Nishiyama, S., et al. 2006, The distance of the Galactic centre derived from infrared photometry of Bulge clump stars, *ApJ*, 647, 1093–1098
- Paczynski, B., & Stanek, K. Z. 1998, Galactocentric distance with the OGLE and *Hipparcos* red clump stars, *ApJ*, 494, L219–L222
- Panagia, N., et al. 1991, Properties of the SN1987A circumstellar ring and the distance to the large Magellanic Cloud, *ApJ*, 380, L23–L26; Erratum, 1992, *ApJ*, 386, L31–L32
- Panagia, N. 2005, A geometric determination of the distance to SN1987A and the LMC, in, *Cosmic Explosions*, ed. J.-M. Marcaide & K. W. Weiler (Berlin: Springer) 585–592
- Perryman, M., et al. 1998, The Hyades: distance, structure, dynamics and age, *A&A*, 331, 81–120
- Perryman, M. 2003, The GAIA mission, in *ASP Conf. Ser. 298, GAIA Spectroscopy, Science and Technology*, ed. U. Munari, 3–12
- Perryman, M. 2009, *Astronomical Applications of Astrometry* (Cambridge: Cambridge University Press)
- Persson, S. E., et al. 2004, New Cepheid period-luminosity relations for the Large Magellanic Cloud: 92 near-infrared light curves, *AJ*, 128, 2239–2264
- Pietrzyński, G., Gieren, W., & Udalski, A. 2003, The Araucaria project. Dependence of mean K , J and I absolute magnitudes of red clump stars on metallicity and age, *AJ*, 125, 2494–2501
- Pietrzyński, G., et al. 2009a, The Araucaria project. Infrared tip of red giant branch distances to the Carina and Fornax dwarf spheroidal galaxies, *AJ*, 138, 459–465
- Pietrzyński, G., et al. 2009b, The Araucaria project. Determination of the Large Magellanic Cloud distance from late type eclipsing binary systems. I. OGLE-051019.64-685812.3, *ApJ*, 697, 862–866
- Pont, F., Mayor, M., & Burki, G., et al. 1994, New radial velocities for classical Cepheids. Local Galactic rotation revisited, *A&A*, 285, 415–439
- Pont, F. 1999, The Cepheid distance scale after Hipparcos, in *ASP Conf. Ser. 167, Harmonizing Cosmic Distance Scales in a Post-Hipparcos Era*, eds. D. Egret, A. Heck, 113–128
- Pont, F., et al. 2004, The chemical enrichment history of the Fornax dwarf spheroidal galaxy from the infrared calcium triplet, *AJ*, 127, 840–860
- Popowski, P., & Gould, G. 1998a, Systematics of RR Lyrae statistical parallax. I. Mathematics, *ApJ*, 506, 259–270
- Popowski, P., & Gould, A. 1998b, Systematics of RR Lyrae statistical parallax. II. Proper motions and radial velocities, *ApJ*, 506, 271–280
- Popowski, P., & Gould, A. 1998c, Systematics of RR Lyrae statistical parallax. III. Apparent magnitudes and extinctions, *ApJ*, 508, 844–853
- Poretti, E., et al. 2008, Variable stars in the Fornax dSph galaxy. II. Pulsating stars below the horizontal branch, *ApJ*, 685, 947–957
- Pritzl, B. J., et al. 2003, *Hubble Space Telescope* snapshot study of variable stars in globular clusters. The inner region of NGC6441, *AJ*, 126, 1381–1401
- Reid, M. J., et al. 2009a, Trigonometric parallaxes of massive star-forming regions. VI. Galactic structure, fundamental parameters, and non-circular motions, *ApJ*, 700, 137–148
- Reid, M. J., et al. 2009b, A trigonometrical parallax of Sgr B2, *ApJ*, 705, 1548–1553
- Rejkuba, M. 2004, The distance to the giant elliptical galaxy NGC 5128, *A&A*, 413, 903–912
- Rejkuba, M., Minniti, D., & Silva, D. R. 2003, Long period variables in NGC5128. I. Catalogue, *A&A*, 406, 75–85
- Rich, R. M., & Origlia, L. 2005, The first detailed abundances for M giants in Baade’s window from infrared spectroscopy, *ApJ*, 634, 1293–1299
- Riess, A. G., et al. 2009, A redetermination of the Hubble constant with the *HubbleSpaceTelescope* from a differential distance ladder, *ApJ*, 699, 539–563
- Rizzi, L. et al. 2007a, Tip of the red giant branch distances. II. Zero-point calibration, *ApJ*, 661, 815–829
- Rizzi, L., et al. 2007b, The distance of the Fornax dwarf spheroidal galaxy, *MN*, 380, 1255–1260
- Saha, A., et al. 2006, Cepheid distances to SNIa host galaxies based on a revised photometric zero point of the *HST* WFPC2 and new PL relations and metallicity corrections, *ApJSup*, 165, 108–137
- Sakai, S., Madore, B. F., & Freedman, W. L. 1996, Tip of the red giant branch distances to Galaxies III, *ApJ*, 461, 713–723

- Sakai, S., Zaritsky, D., & Kennicutt, R. C. 2000, The tip of the red giant branch distance to the Large Magellanic Cloud, *AJ*, 119, 1197–1204
- Salaris, M., & Girardi, L. 2002, Population effects on the red giant clump absolute magnitude: the K band, *MN*, 337, 332–340
- Salaris, M., & Girardi, L. 2005, Tip of the red giant branch distances to galaxies with composite stellar populations, *MN*, 357, 669–678
- Salaris, M., et al. 2001, On the white dwarf distances to galactic globular clusters, *A&A*, 371, 921–931
- Salaris, M., Cassisi, S., & Weiss, A. 2002, Red giant branch stars: The theoretical framework, *PASP*, 114, 375–402
- Sale, S. E., et al. 2010, The structure of the outer Galactic disc as revealed by IPHAS early A stars, *MN*, 402, 713–723
- Sandage, A. 1958, Current problems in the extragalactic scale, *ApJ*, 127, 513–526
- Sandage, A., & Tammann, G. A. 1969, The double Cepheid CE Cassiopeiae in NGC 7790: Tests of the theory of the instability strip and the calibration of the period-luminosity relation, *ApJ*, 157, 683–708
- Sandage, A. 1972, Classical Cepheids: Cornerstones to extragalactic distances? *QJRAS*, 13, 202–221
- Sandage, A. 1990, The vertical height of the horizontal branch - the range in the absolute magnitudes of RR Lyrae stars in a given globular clusters, *ApJ*, 350, 603–630
- Sandage, A. 1994a, Bias properties of extragalactic distance indicators. I, *ApJ*, 430, 1–12
- Sandage, A. 1994b, Bias properties of extragalactic distance indicators. II, *ApJ*, 430, 13–28
- Sandage, A., & Tammann, G. A. 2006, Absolute magnitude calibrations of populations I and II Cepheids and other pulsating variables in the instability strip of the Hertzsprung-Russell diagram, *ARAA*, 44, 93–140
- Saviane, I., Held, E. V., & Bertelli, G. 2000, The stellar populations of the Fornax dwarf spheroidal galaxy, *A&A*, 355, 56–68
- Schechter, P. L. 1980, Mass-to-light ratios for elliptical galaxies, *AJ*, 85, 801–811
- Schmidt-Kaler, Th. 1982, Physical parameters of the stars, in *Landolt-Börnstein New series Group VI, Vol 2b, 1*, eds. K. Schaifers, H. H. Voigt (Berlin: Springer-Verlag)
- Shapley, H. 1938, Two stellar systems of a new kind, *Nature*, 142, 715–716
- Siegel, M. H., et al. 2002, Star counts redivivus. IV. Density laws through photometric parallaxes, *ApJ*, 578, 151–175
- Smith, H. A. 1995, *RR Lyrae stars* (Cambridge: Cambridge University Press)
- Smith, H. 2003, Is there really a Lutz-Kelker bias? Reconsidering calibration with trigonometrical parallaxes, *MN*, 338, 891–902
- Smith, H. 2006, Beware λ -truncation! Sample truncation and bias in luminosity calibration using trigonometric parallaxes, *MN*, 365, 469–476
- Sollima, A., Cacciari, C., & Valenti, E. 2006, The RR Lyrae period-K luminosity relation for globular clusters; an observational approach, *MN*, 372, 1675–1680
- Soszyński, I., et al. 2003, The optical gravitational lensing experiment, catalog of RR Lyrae variables in the Large Magellanic Cloud, *AcA*, 53, 93–116
- Soszyński, I., et al. 2008, The optical gravitational lensing experiment. The OGLE-III Catalog of Variable Stars II, *AcA*, 58, 293–312
- Spaenhauer, A. M. 1978, A systematic comparison of four methods to derive stellar space densities, *AA*, 65, 313–321
- Sterken, C., & Jaschek, C. 1996, *Light Curves of Variable Stars* (Cambridge: Cambridge University Press)
- Strand, K. Aa. 1963, Trigonometric stellar parallaxes, in *Basic Astronomical Data*, ed. K. Aa. Strand (Chicago: University of Chicago Press), 55–63
- Szewczyk, O., et al. 2008, The Araucaria project. The distance of the Large Magellanic Cloud from near-infrared photometry of RR Lyrae variables, *AJ*, 136, 272–279
- Tabur, V., Kiss, L. L., & Bedding, T. R. 2009, Hipparcos calibration of the tip of the red giant branch, *ApJ*, 703, L73–L75
- Teerikorpi, P. 1997, Observational selection bias affecting the determination of the extragalactic distance scale, *ARAA*, 35, 101–136
- Tolstoy, E., et al. 2003, VLT/EVES abundances in four nearby dwarf spheroidal galaxies. II. Implications for understanding galaxy evolution, *AJ*, 125, 707–726
- Udalski, A., Kubiak, M., & Szymanski, M., 1997, Optical gravitational lensing experiment OGLE2, *AcA*, 47, 319–344
- Udalski, A., et al. 1999, The optical gravitational lensing experiment. Cepheids in the Magellanic Clouds, *AcA*, 49, 201–221
- van den Bergh, A. 1975, The extragalactic distance scale, in *Galaxies and the Universe*, eds. A. Sandage et al. (Chicago: University of Chicago Press), 509–539
- van der Kruit, P. C. 2000, The Milky Way compared to other galaxies, in *The legacy of J C Kapteyn*, eds. P. C. van der Kruit, K. van Berkel (Dordrecht: Kluwer), 229–323

- van Langevelde, H. J., van der Heiden, R., & van Schooneveld, C. 1990, Phase lags from multiple flux curves of OH/IR stars, *A&A*, 239, 193–204
- van Leeuwen, F. 2007, *Hipparcos, the new reduction of the raw data* (Berlin: Springer)
- van Leeuwen, F., et al. 1997, First results from Hipparcos trigonometrical parallaxes of Mira-type variables, *MN*, 287, 955–960
- van Leeuwen, et al. 2007, Cepheid parallaxes and the Hubble constant, *MN*, 379, 723–737
- van Leeuwen, F. 2009, Parallaxes and proper motions for 20 open clusters as based on the new Hipparcos catalogue, *A&A*, 497, 209–242
- Vlemmings, W. H. T., et al. 2003, VLBI astrometry of circumstellar OH masers: Proper motions and parallaxes for four AGB stars, *A&A*, 407, 213–224
- Vlemmings, W. H. T., & van Langeveldt, H. J. 2007, Improved VLBI astrometry of OH maser stars, *A&A*, 472, 547–553
- Wallerstein, G. et al. 1992, Metallic-line and H α radial velocities of seven southern Cepheids – A comparative analysis, *MN*, 259, 474–488
- Wegner, W. 2006, Absolute magnitudes of OB and Be stars based on Hipparcos parallaxes - II, *MN*, 371, 185–192
- Wegner, W. 2007, Absolute magnitudes of OB and Be stars based on Hipparcos parallaxes - III [Note. This title is erroneous], *MN*, 374, 1549–1556
- Whitelock, P. A., & Catchpole, R. M. 1992, The shape of the Bulge from IRAS Miras, in *The Center, Bulge and Disc of the Milky Way*, ed. L. Blitz (Dordrecht: Kluwer), 103–110
- Whitelock, P. A., & Feast, M. W. 2000, Hipparcos parallaxes for Mira-like long period variables, *MN*, 319, 759–770
- Whitelock, P. A., Feast, M. W., & Catchpole, R. M. 1991, IRAS sources and the nature of the Galactic Bulge, *MN*, 248, 276–312
- Whitelock, et al. 2003, Obscured asymptotic giant branch variables in the Large Magellanic Cloud and the period-luminosity relation, *MN*, 342, 86–104
- Whitelock, P. A., Feast, M. W., & van Leeuwen, F., 2008, AGB variables and the Mira period-luminosity relation, *MN*, 386, 313–323
- Whitelock, P. A., et al. 2009, Asymptotic giant branch stars in the Fornax dwarf spheroidal galaxy, *MN*, 394, 795–809
- Wood, P. R. 2000, Variable red giants in the LMC, Pulsating stars and binaries? *PASA*, 17, 18–21
- Wood, P. R., Bessell, M. S., & Fox, M. W. 1983, Long period variables in the Magellanic Clouds - supergiants, AGB stars, supernovae precursors, planetary nebulae precursors, enrichment of the interstellar medium, *ApJ*, 272, 99–115

17 Globular Cluster Dynamical Evolution

Melvyn B. Davies

Lund Observatory, Lund, Sweden

1	<i>Introduction</i>	881
2	<i>Cluster Models</i>	884
2.1	Plummer Model	885
2.2	Isothermal Sphere	886
2.3	King Model	888
2.4	Modeling Clusters on a Computer	891
2.4.1	Full N-Body Codes	893
2.4.2	Fokker-Planck Approach	893
2.4.3	Gas Models	894
2.4.4	Monte Carlo Codes	894
3	<i>Internal Dynamical Evolution</i>	895
3.1	Processes Occurring at the Early Stages	895
3.1.1	Phase Mixing and Violent Relaxation	895
3.1.2	Mass Loss from the Cluster	897
3.2	Two-Body Relaxation	898
3.3	Mass Segregation	900
3.3.1	Observations of Mass Segregation	901
3.4	Core Evolution	901
3.4.1	The Gravothermal Catastrophe	901
3.4.2	Core Collapse	902
3.4.3	Gravothermal Oscillations	903
4	<i>Complications and Additional Effects</i>	904
4.1	Stellar Collisions	905
4.2	Stellar Binaries	906
4.2.1	Making Binaries via Tidal Capture	906
4.2.2	Binary/Single Encounter Dynamics and Heating	907
4.2.3	The Primordial Binary Population	908
4.3	Black Holes	909
4.3.1	Stellar-Mass Black Holes	909
4.3.2	Observational Evidence for Intermediate-Mass Black Holes	911
4.3.3	Producing Intermediate-Mass Black Holes in Clusters	912
4.4	Multiple Stellar Populations	913

5	<i>Cluster Survival</i>	914
5.1	Stellar Evaporation and Tidal Truncation	915
5.2	Disk and Bulge Shocking	915
5.3	Inspiral Due to Dynamical Friction	916
5.4	The Combined Effect on the Cluster Population	917
	<i>Acknowledgments</i>	919
	<i>References</i>	920

Abstract: Globular clusters are some of the oldest structures in the universe. They typically contain 10^5 – 10^6 stars and thus they are excellent laboratories in which to explore the effects of dynamical evolution on self-gravitating systems. Two-body scattering between stars is responsible for energy transport within the cluster. Through this mechanism energy flows from stars in the cluster center to stars in the cluster halo. Because a self-gravitating system has a negative heat capacity, when the stars in the cluster core lose energy via two-body scattering, the core contracts and heats up (in the sense that the stellar velocities increase). The flow of energy is then accelerated leading to extremely high central densities; a process known as core collapse. The time taken to reach this stage is a function of the relaxation timescale within the cluster (the timescale required for an accumulated number of distant two-body scatterings to be effective). This timescale is a function of the properties of the cluster mass and radius, and thus clusters observed today have a broad range of central densities. Energy input from binaries within the cluster core can halt core collapse. The cluster will then expand, with energy input continuing to come from binaries in the core. For clusters having a sufficiently large number of stars, the expansion can be unstable leading to so-called gravothermal oscillations. Dynamical processes within the cluster might also produce exotic objects such as intermediate-mass black holes. Clusters do not exist in isolation. The galaxy affects the evolution of a cluster via three mechanisms: (1) dynamical friction will cause clusters at relatively small galactocentric radii to spiral into the galactic center, (2) the galactic tidal field will strip stars from the outer regions of stellar clusters, and (3) the entire cluster will be stirred up as it passes through the galactic disk or close to the bulge which will enhance cluster mass loss. Thus the clusters seen today may only be a small subset of the original population.

Keywords: Core collapse, Fokker–Planck method, Globular cluster, Gravothermal catastrophe, Isothermal sphere, King model, Mass loss, Mass segregation, Plummer model, Stellar collisions, Stellar evaporation, Tidal capture, Tidal truncation, Two-body relaxation, Violent relaxation

1 Introduction

This chapter concerns the dynamical evolution of globular clusters. There are about 150 globular clusters in our galaxy, found in the bulge and halo. They are all rather old. They hold the interest of astronomers for several reasons. They are some of the oldest structures in the universe, containing typically 10^5 – 10^6 stars, and are relatively simple objects (compared to an entire galaxy). They can be used to study stellar evolution by observing their stellar populations. They are sites for the production of various exotic stellar objects. They are also very productive laboratories in which to study the processes at play within self-gravitating systems. The latter will be focussed on here, considering the distribution of stars within a cluster, and how this distribution changes with time, due to processes occurring within the cluster (such as stars being deflected or scattered by each other) as well as those processes caused by the external influence of our galaxy. This chapter will not discuss in detail the production of exotic objects such as millisecond pulsars and ultratight binaries containing black holes, white dwarfs, or neutron stars. Populations of objects observed in globular clusters are dealt with in the chapter on the contents of globular clusters.

In order to consider the dynamical evolution of globular clusters, one needs to understand how to first calculate static cluster models (where the properties such as the density distribution are constant with time). The internal and external processes which cause a cluster to change must be considered, for example, with clusters becoming more or less centrally concentrated or losing stars via tidal stripping. It will turn out that the processes which drives the cluster evolution operate on timescales which are much longer than the time it takes stars to cross the cluster (the so-called crossing time). Thus beginning with static models is reasonable. Theoretical calculations must also be connected with the available observational data in order to understand how observations may help constrain our cluster models. Our task is also made more complex by the myriad of interactions possible within the stellar population contained in the cluster. Stars are neither static – they evolve – nor are they point masses. A large fraction of them can also be in multiple systems, such as binaries and triples.

The key concepts and ideas, which will be discussed in more detail in later sections of this chapter, are first outlined below:

(a) Distribution functions and their use in making cluster models:

When dealing with a self-gravitating cluster of stars, the distribution of stars must be considered, not only in the three spatial dimensions, but also in the three dimensions of their velocities. By considering how the stars are distributed in this six-dimensional phase space (known as the distribution function), one is able to compute the density distribution within a cluster, and also the gravitational potential. It turns out that there are choices of distribution function which provide good fits to the observed clusters while being both relatively simple in form as well as physically motivated.

(b) The role and consequences of two-body scattering:

When two stars within a cluster pass relatively close to each other they will be deflected by their mutual gravitational attractive force. Kinetic energy may also be transferred from one star to the other. Over time, there is an accumulated effect of many such scatterings, with the stars being driven toward equipartition, in other words they tend to share equal amounts of kinetic energy at the same location within a cluster. This is known as two-body relaxation and it is the engine which drives the dynamical evolution of stellar clusters. Through its mediation, heavier stars will sink into the cluster core, while lighter ones will evaporate from the cluster. As will be seen in later sections, the flow of energy from stars in the center to those in the halo leads to cluster core collapse, where cluster central densities reach very high values. The time taken to reach core collapse for some clusters is in the range 10^9 – 10^{10} years while for others it is much longer. Thus one would expect to see clusters in various stages of their dynamical evolution, i.e., some more centrally concentrated than others. This is indeed observed.

(c) What observations reveal about stellar clusters:

The past two decades have seen a wealth of observations of stellar clusters, for a number of reasons. The high-resolution capabilities of HST has meant that individual stars can be imaged even in crowded cluster cores. The stellar population can then be studied via photometry. Surface luminosity profiles can also be used to fit to theoretical cluster models. Spectroscopic studies provide important information about the velocity dispersions in clusters which help theorists to constrain cluster models. Observations at other wavelengths, particularly radio and X-ray, also afford us different views of clusters: one is then

able to perform an audit of the population of millisecond radio pulsars and X-ray binaries in clusters. These in turn provide us with important clues concerning the dynamical history of stellar clusters.

- (d) How stellar clusters can be modeled on a computer:

Modeling the dynamical evolution of a stellar cluster on a computer turns out to be rather difficult. Firstly, there are an enormous range of timescales within the problem: stars cross the cluster in 10^5 – 10^6 years or so while the global properties of the cluster typically evolve on a timescale of 10^9 – 10^{10} years or longer. A cluster might also contain tight binary star systems with orbital periods of only hours. One must also take care to integrate the motion of the stars accurately: it is the accumulation of small defelections via two-body scattering which drive the cluster evolution. The past two decades has seen great improvements in both computer power but also numerical approaches to modeling stellar cluster evolution. Direct N-body techniques can now be applied, for lower-mass clusters at least, on custom computer hardware (such as computer graphics cards). There are also a number of other techniques which provide credible approximations to a real cluster together with the benefit of being much faster than direct N-body codes.

- (e) Stellar cluster ecology and how it adds complexity:

The evolution of a stellar cluster treating it as a self-gravitating ensemble of single point masses is first considered. In reality, many stars may be in binary systems, which can evolve with mass flowing between the two stars, or which encounter other binaries or single stars. The high number densities of stars in the cluster cores implies that stars may have physical collisions with other stars. This provides channels for exotica production but also can affect the dynamical evolution of the system (for example, as physical collisions may lead to mass loss from the cluster). Stellar-mass black holes, produced in a subset of core-collapse supernovae, may be retained in at least some globular clusters. They may sink (as they are heavier than most other stars in the cluster) forming their own separate core in the cluster center. This population of black holes may provide energy to the cluster, causing it to expand. It could be that in some circumstances an intermediate-mass black hole is formed via the merger of several stellar-mass black holes. Increasing observational evidence suggests that the population of stars in a globular cluster is not simple: many clusters show evidence for multiple episodes of star formation. This challenges the simple picture where all stars in the cluster were formed in one single event.

- (f) How the cluster population is shaped by our galaxy:

An isolated cluster will evolve without any outside influence. However the globular clusters one observes in our galaxy have been, at least in part, shaped by its external influence. The tidal field of the galactic potential will truncate the clusters, removing stars which wander further out. The stars within a cluster will be stirred up as the cluster passes through the galactic disk or passes close to the bulge on its orbit within the galaxy. A process known as dynamical friction will act as a drag force on globular clusters. Those forming relatively close to the galactic center may spiral in and be lost. It turns out that a combination of all these processes will contribute to the modification of the globular cluster population. It is unlikely that all globular clusters formed in our galaxy have survived until today: a significant fraction has been destroyed by these three processes.

In [Sect. 2](#) the construction of static cluster models is reviewed, beginning with a discussion of distribution functions. Various cluster models are then described: Plummer models,

the isothermal sphere, King models, and King–Michie models. The various possible numerical approaches to modeling cluster evolution is also presented. The internal dynamical evolution of a cluster is considered in [Sect. 3](#). Beginning with the processes occurring very early on in the cluster’s history, two-body relaxation is described and the timescale for it to occur is derived. The effects of two-body relaxation – mass segregation, core collapse, and bounce – are then explained. [Section 4](#) considers the process of stellar collisions together with the effects of a population of stellar binaries. The role played by stellar-mass black holes is also discussed. The observational evidence for multiple stellar populations within clusters is briefly reviewed. In [Sect. 5](#) the evolution of the cluster population is considered allowing for the destructive effects of our galaxy in the form of tidal stripping, disk and bulge shocking, and dynamical friction.

There are a great number of review articles available concerning both dynamical evolution of self-gravitating systems and globular clusters. Several of these review articles are cited throughout this chapter. Here the reviews by Hut et al. (1992a), Meylan and Heggie (1997), and Portegies Zwart et al. (2010) are noted. There are also a number of useful books and resources available on the web. Books on this topic include: Spitzer (1987), Ashman and Zepf (1998), Heggie and Hut (2003), Aarseth (2003), Binney and Tremaine (2008), and Aarseth et al. (2008). Of the many useful websites, two are identified here. MODEST is an international collaboration of those working on modeling dense stellar systems (www.manybody.org). William Harris maintains an online catalogue of globular cluster parameters for the globulars in our galaxy (<http://www.physics.mcmaster.ca/Globular.html>).

2 Cluster Models

Static models for globular clusters are considered first where the mass density of stars and the gravitational potential are independent of time. As will be seen in later sections, in reality clusters are never static in this sense, for example, two-body scattering will drive energy flow and dynamical evolution in a cluster but often on very long timescales. Static models have their uses, however, and real clusters can be modeled by them.

In order to make models for stellar clusters, a number of assumptions are made. Firstly, it is assumed that the granularity of the cluster may be ignored. In other words, the sea of stars may be considered as a continuum which results in a total, smooth, gravitational potential well. The force felt by any individual star is then proportional to the gradient of this potential.

Consider first the distribution function $f(\mathbf{r}, \mathbf{v})$, which is a function of position \mathbf{r} and velocity \mathbf{v} . The number of stars contained in a volume element $d\mathbf{r}$ centered on a position \mathbf{r} , having velocities contained in the element $d\mathbf{v}$ centered around velocity \mathbf{v} at a given time, is given by

$$dn(\mathbf{r}, \mathbf{v}) = f(\mathbf{r}, \mathbf{v}) d\mathbf{r} d\mathbf{v} \quad (17.1)$$

One may integrate over all velocities to compute the number density of stars as a function of position

$$n(\mathbf{r}) = \int f(\mathbf{r}, \mathbf{v}) d\mathbf{v} \quad (17.2)$$

The gravitational potential is related to the density via Poisson's equation

$$\nabla^2 \phi(\mathbf{r}) = 4\pi G \rho(\mathbf{r}) \quad (17.3)$$

where the mass density of stars $\rho(\mathbf{r}) = mn(\mathbf{r})$, where m is the mass of an individual star if all stars have the same mass, or more generally $\rho(\mathbf{r}) = \sum m_i n_i(\mathbf{r})$ where the stellar population is split up into various mass bins. In the following sections it will be further assumed that the potential ϕ is a function only of spherical radius, r , and that the distribution function f is a function only of radius r , radial velocity v_r , and transverse velocity v_t . The energy per unit mass E and angular momentum per unit mass J are both constant for a star on a given orbit, and are given by

$$E = \frac{v^2}{2} + \phi(r) \quad (17.4)$$

$$J = rv_t = rv \sin \theta \quad (17.5)$$

where θ is the angle between \mathbf{r} and \mathbf{v} . Rather than expressing the distribution function in terms of position and velocity, it will turn out to be more convenient to consider it as a function of E and J . It should be recalled that the exact value of the potential ϕ includes an arbitrary constant; one is therefore free to set the location of the zero-point of the potential.

2.1 Plummer Model

A class of stellar models can be obtained by assuming an isotropic velocity distribution throughout the cluster, i.e., the distribution function can be written as a function of energy only

$$f \propto (-E)^p \quad (17.6)$$

for $E < 0$ or $f = 0$ otherwise. As seen above, the specific energy $E = v^2/2 + \phi(r)$, where the potential is taken to be $\phi(r) = 0$ at the cluster surface. [Equations \(17.2\)](#) and [\(17.3\)](#) can now be used to compute the stellar mass density as a function of cluster radius. For $p > -1$ one obtains (Spitzer 1987)

$$\rho(r) \propto [-\phi(r)]^{p+3/2} \quad (17.7)$$

Polytropes follow the above equation where the polytropic index n is related to p via the equation $n = p + 3/2$. Analytic solutions are available for polytropes of index $n = 0$, $n = 1$, and $n = 5$ (e.g., Chandrasekhar 1967). It turns out that a polytrope of index $n = 5$ provides a reasonable fit to many stellar clusters, with the density showing a somewhat flattened core and dropping off to smaller values at larger radii. Plummer first used it to fit to stellar clusters, hence it has become known as the Plummer model (Plummer 1911). Using a Plummer model for a cluster is very appealing as, because of its analytic nature, many properties of the cluster have simple algebraic expressions (for an extensive table see Heggie and Hut (2003)). For example, the three-dimensional mass density for a polytrope of index $n = 5$ is given by

$$\rho(r) = \frac{3M}{4\pi a^3} \left(1 + \frac{r^2}{a^2}\right)^{-5/2} \quad (17.8)$$

where M is the total cluster mass and a is a constant connected with the length scale of the cluster. By integrating the expression for the density, one may obtain the expression for the enclosed mass as a function of radius

$$M(r) = M \left(1 + \frac{a^2}{r^2} \right)^{-3/2} \quad (17.9)$$

The gravitational potential is given by

$$\phi(r) = -\frac{GM}{a} \left(1 + \frac{r^2}{a^2} \right)^{-1/2} \quad (17.10)$$

The projected mass density, Σ , is also of interest as it is connected with the surface brightness of a cluster, which is a quantity which is frequently measured in observed clusters.

$$\Sigma(d) = \frac{M}{\pi a^2} \left(1 + \frac{d^2}{a^2} \right)^{-2} \quad (17.11)$$

where d is now the projected distance from the cluster center. The observed core radius r_c is defined as being the projected distance from the cluster center where the surface brightness drops to half its central value. In other words $(1 + r_c^2/a^2)^{-2} = 1/2$. Rearrangement yields that $r_c = \sqrt{(2^{1/2} - 1)}a \simeq 0.6436a$.

The generalization to the surface density given above is provided by Elson et al. (1987) who fitted surface brightness profiles to observed young clusters having the form

$$\Sigma(d) = \Sigma_0 \left(1 + \frac{d^2}{a^2} \right)^{-\gamma/2} \quad (17.12)$$

and found that young clusters in the LMC and SMC could be fit with $\gamma \simeq 3.5 - 4$.

2.2 Isothermal Sphere

It has been seen in the previous section how the distribution for a polytrope of index $n = 5$ could be used to model a distribution of stars. Here the situation as $n \rightarrow \infty$ is considered. In other words the stellar cluster is equivalent to a polytropic gas where $\gamma = 1$. Such a gas has an isothermal equation of state (i.e., $p = K\rho$). In order to determine the density distribution of a sphere of such a gas, one can take the Lane-Emden equation and consider the limit as $n \rightarrow \infty$ (see Hunter 2001). An alternative derivation is given in Binney and Tremaine (2008) and is outlined below. Suppose that the stellar cluster has a distribution function that follows a Maxwellian, i.e.,

$$f(E) = K e^{-E/\sigma^2} = K e^{-(v^2/2 + \phi(r))/\sigma^2} \quad (17.13)$$

where K is a suitable constant and σ is the velocity dispersion. One can integrate over all velocities to obtain an expression for the density of stars

$$\rho(r) = \rho_1 e^{-\phi(r)/\sigma^2} \quad (17.14)$$

where ρ_1 is a constant. Plugging this expression into Poisson's equation, one arrives at the following expression:

$$\frac{d}{dr} \left(r^2 \frac{d \ln \rho}{dr} \right) = -\frac{4\pi G}{\sigma^2} r^2 \rho \quad (17.15)$$

One can reach a similar equation by considering the equation of hydrostatic equilibrium for gas having an isothermal equation of state, where the pressure $p = k_B \rho T/m$. The two equations are the same if $\sigma^2 = k_B T/m$. In other words, the density distribution of a stellar cluster, where the stars have the distribution function given by (● 17.13), is the same as for an isothermal self-gravitating sphere of gas.

There is one analytic solution to (● 17.15). By setting $\rho = Kr^{-\gamma}$, where K and γ are constants, and plugging into (● 17.15), one obtains

$$-\gamma = -\left(\frac{4\pi G}{\sigma^2} \right) Kr^{2-\gamma} \quad (17.16)$$

hence one sees that $\gamma = 2$ and $K = \sigma^2/(2\pi G)$. Thus the density is given by

$$\rho(r) = \frac{\sigma^2}{2\pi G r^2} \quad (17.17)$$

The above solution is known as the singular isothermal sphere. One can obtain the enclosed mass $M(r)$ as a function of radius r ,

$$M(r) = \frac{2\sigma^2 r}{G} \quad (17.18)$$

while the surface density is given by

$$\Sigma(d) = \frac{\sigma^2}{2Gd} \quad (17.19)$$

where d is the projected distance from the cluster center. In addition, the circular orbital speed $v_c(r) = \sqrt{2}\sigma$ and the potential $\phi(r) = 2\sigma^2 \ln(r) + K$, where K is some constant.

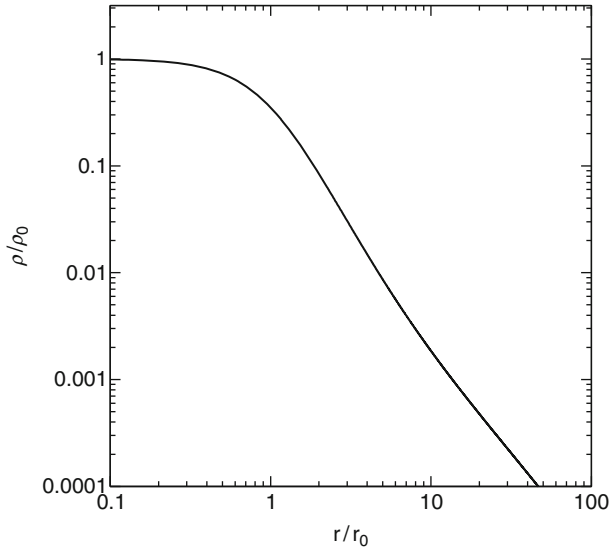
A major problem with the singular isothermal sphere is that it has infinite density for $r = 0$. One can, however, obtain a second solution to (● 17.15) which is well-behaved at the origin. It turns out to be more convenient to define new variables $\tilde{\rho} = \rho/\rho_0$ and $\tilde{r} = r/r_0$ where ρ_0 is the central density and r_0 is known as the King radius and is given by

$$r_0 = \sqrt{\frac{9\sigma^2}{4\pi G \rho_0}} \quad (17.20)$$

In terms of $\tilde{\rho}$ and \tilde{r} (● 17.15) becomes

$$\frac{d}{d\tilde{r}} \left(\tilde{r}^2 \frac{d \ln \tilde{\rho}}{d\tilde{r}} \right) = -9\tilde{r}^2 \tilde{\rho} \quad (17.21)$$

This equation can then be solved numerically, by integrating outward from $\tilde{r} = 0$. The density is plotted as a function of radius in ● Fig. 17-1. The choice of the definition of r_0 is particularly useful as it turns out that the surface density drops to very close to half its central value at a projected separation of r_0 and is therefore analogous to the observational definition of core radius which is the distance at which the surface brightness drops to half its central value.



■ Fig. 17-1
The density distribution (ρ/ρ_0) of an isothermal sphere

2.3 King Model

A problem with the isothermal sphere is that the enclosed mass is unlimited (as density is proportional to r^{-2} , mass keeps on increasing for increasing radii). A distribution function is required that is similar to the isothermal sphere close in (i.e., deep inside the cluster where the stars are deeply bound) but which drops off more quickly further out and drops to zero at an edge.

A simple modification to the Maxwellian distribution function given in (● 17.13) is the so-called lowered Maxwellian where the tail of the distribution is removed. The new distribution function is given by (e.g., Spitzer 1987)

$$f(E) = K(e^{-E/\sigma^2} - e^{-E_c/\sigma^2}) \quad (17.22)$$

for $E < E_c$, or $f(E) = 0$ if $E > E_c$. The idea here is that the stellar cluster sits inside the galactic tidal field which will remove stars from the cluster exterior to some tidal radius. Stellar cluster models having a distribution function of the form shown above have become known as King models after King (1966) who made them well known.

Following the discussion in Binney and Tremaine (2008), one may rewrite the distribution function in terms of the relative potential $\psi = -\phi + \phi_0$, where ϕ_0 is a suitably chosen constant, and the relative energy $\mathcal{E} = \psi - v^2/2$. One is free to choose ϕ_0 such that $f(\mathcal{E}) > 0$ for $\mathcal{E} > 0$ and $f(\mathcal{E}) = 0$ for $\mathcal{E} < 0$. Thus (● 17.22) becomes (Binney and Tremaine 2008)

$$f(\mathcal{E}) = \rho_1(2\pi\sigma^2)^{-3/2}(e^{\mathcal{E}/\sigma^2} - 1) \quad (17.23)$$

for $\mathcal{E} > 0$ or $f(\mathcal{E}) = 0$ otherwise. This distribution function may then be integrated over all velocities to obtain an expression for the density as a function of ψ ((4.111) from Binney and Tremaine 2008)

$$\begin{aligned}\rho(\psi) &= \frac{4\pi\rho_1}{(2\pi\sigma^2)^{3/2}} \int_0^{\sqrt{2\psi}} \left[\exp\left(\frac{\psi - v^2/2}{\sigma^2}\right) - 1 \right] v^2 dv \\ &= \rho_1 \left[e^{\psi/\sigma^2} \operatorname{erf}\left(\frac{\sqrt{\psi}}{\sigma}\right) - \sqrt{\frac{4\psi}{\pi\sigma^2}} \left(1 + \frac{2\psi}{3\sigma^2}\right) \right]\end{aligned}\quad (17.24)$$

where erf, the so-called error function, is given by $\operatorname{erf}(X) = (2/\sqrt{\pi}) \int_0^X e^{-x^2} dx$. Poisson's equation for this spherically symmetric system is given by

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d\psi}{dr} \right) = -4\pi G \rho(\psi) \quad (17.25)$$

Hence one may use (17.24) and (17.25) to integrate $\psi(r)$ numerically outward from $r = 0$, and thus finding $\rho(r)$. As the edge of the cluster is approached, ψ approaches zero, as does the range of speeds the stars may have in the integral in (17.24), and thus the density drops quickly to zero. The radius at which the cluster density drops to zero is typically labeled r_t , for tidal radius. The concentration of a stellar cluster is defined by

$$c = \log_{10}(r_t/r_0) \quad (17.26)$$

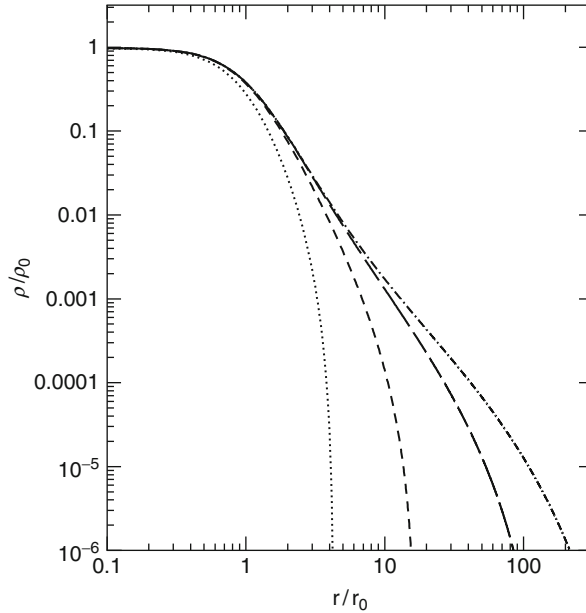
where r_t is the tidal radius. The concentration is a function of the depth of the potential well of the cluster, $W_0 = \psi(0)/\sigma^2$. Larger values of W_0 produce more concentrated clusters. This can be seen in Fig. 17-2 where the density profiles for King models where $W_0 = 3, 6, 9$, and 12 are plotted. As will be seen in Sect. 3, over time clusters will evolve dynamically toward more concentrated configurations.

Two extensions to King models are now considered. It is possible to change the distribution function slightly in order to include velocity anisotropy. In King–Mitchie models, the distribution function has the general form (Michie 1963)

$$f(E, J) = K e^{-\beta J^2} (e^{-E/\sigma^2} - e^{-E_c/\sigma^2}) \quad (17.27)$$

In this case, velocity isotropy is preserved in the cluster center, but the velocity distribution becomes progressively radial at larger radii. Here $\beta = 1/2r_a^2\sigma^2$ where r_a is known as the anisotropy radius. As r_a becomes a large number, the above distribution function returns to the original as given in (17.22).

Thus far, stellar clusters have been considered which contain stars that all have identical masses. In reality, stars have a range of masses. The old globular clusters in our galaxy today contain only low-mass stars on the main sequence (around $0.8 M_\odot$ or less) but they also contain the compact remnants of more massive stars. Stars of intermediate mass produce white dwarfs, which mostly have masses around $0.6 M_\odot$, whereas the most massive stars will have produced neutron stars ($M_{\text{ns}} \simeq 1.2 - 1.4 M_\odot$) or stellar-mass black holes ($M_{\text{bh}} \simeq 3 - 15 M_\odot$).



■ Fig. 17-2

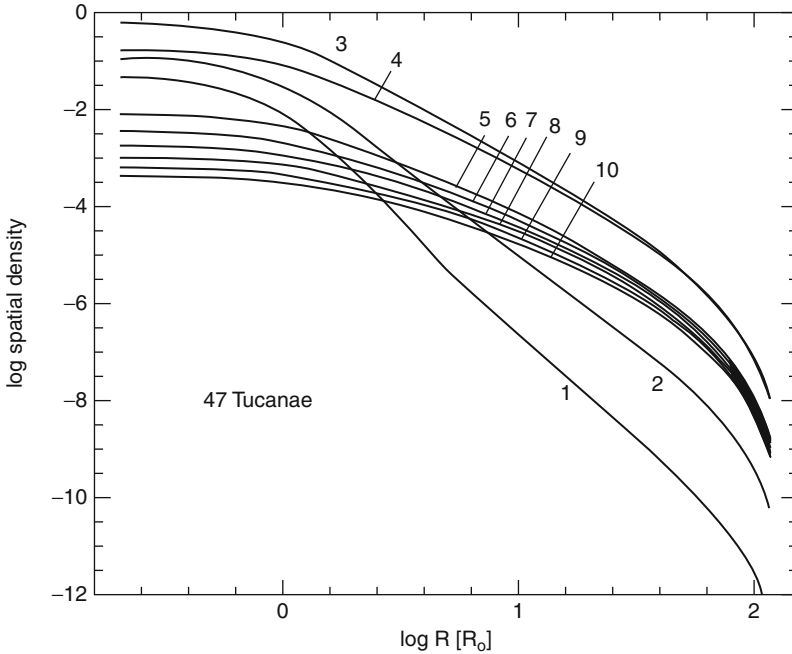
Density as a function of radius for King models with (from left to right) $W_0 = 3, 6, 9,$ and $12,$ where $W_0 = \psi(0)/\sigma^2$. The radius is given in terms of the King radius r_0 which is defined in (17.20)

One can construct a cluster model based around a set of King–Mitchie distribution functions, one for each mass bin of stellar objects, i.e., for the i th mass bin of stars, one obtains (Gunn and Griffin 1979)

$$f_i(E, J) = K e^{-\beta J^2} (e^{-A_i E} - 1) \quad (17.28)$$

The key point here is that scattering between stars will ensure equipartition of energy amongst the stars of different masses, i.e., $m_i \sigma_i^2 = \text{constant}$ at some radius within the cluster. Hence $A_i \propto m_i$. One can then solve as before, integrating outward for a given value of W_0 , assuming the contribution to the central density made by each mass bin, α_i , until one reaches the edge of the cluster. The mass contained in each mass bin for a given set of α_i can then be computed. One adjusts the values of α_i until the fractions of mass contained in the various mass bins agree with the mass function assumed for the entire cluster.

Observationally, one may fit both the surface brightness profile of a cluster and the radial-velocity profile, finding the properties of the multi-mass King–Mitchie model which provides the best fit to the observations. These properties are the depth of the potential well W_0 ; the properties of the stellar population; the contribution each bin makes to the central density, α_i ; and the anisotropy radius which determines the anisotropy of the velocity dispersion within the system (contained in β in (17.28)). King–Mitchie models have been fit to several globular clusters, including ω Centauri (Meylan 1987) and 47 Tucanae (Meylan 1988). In Fig. 17-3, the spatial density contributions for the various mass bins for a model for 47 Tucanae are plotted. One can see from this figure that the most massive mass bins are more centrally concentrated



■ Fig. 17-3

Log spatial density as a function of log radius (in units of the King radius r_0) for a multi-mass King model for 47 Tucanae from Meylan (1988). The lines shown are for stars of different masses. Line 1 being for the most massive stars ($1.4 M_{\odot}$ neutron stars), line 2 for massive white dwarfs, lines 3 and 4 for stars for massive main-sequence stars and white dwarfs, and lines 5–10 for lower-mass main-sequence stars (Fig. 5 from Meylan (1988), reproduced with permission)

and make up a larger fraction of the mass in the center compared to their contribution averaged over the entire cluster.

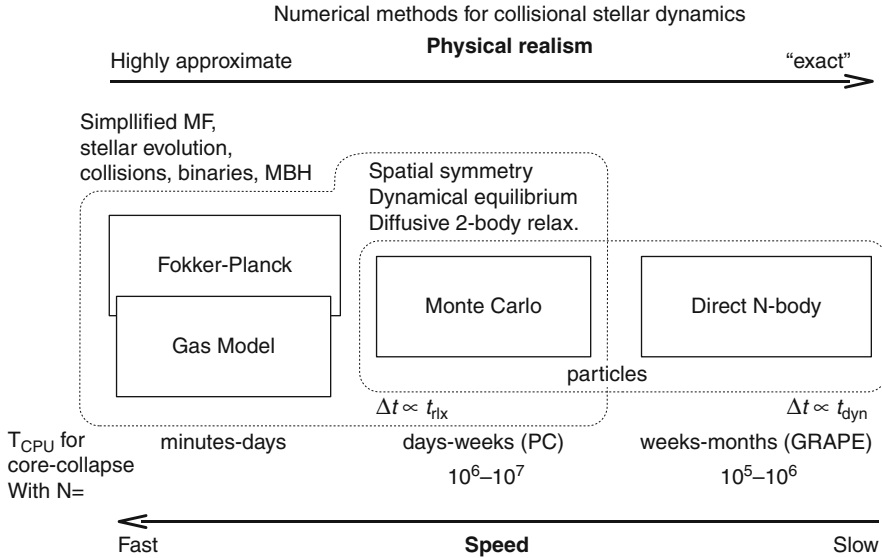
Before this section is closed, the important difference between King models as described here and King profiles will be noted. The following expression for the projected surface density provides a good fit to observed clusters (King 1962)

$$\Sigma(d) = \Sigma_0 \left[\frac{1}{(1 + (d/r_c)^2)^{1/2}} - \frac{1}{(1 + (r_t/r_c)^2)^{1/2}} \right]^2 \tag{17.29}$$

where d is the projected distance to the cluster center, r_c is the core radius, and r_t is the tidal radius of the cluster. It should be emphasized that this simple formula is an empirical fit, and not derived from the distribution function given in (● 17.22).

2.4 Modeling Clusters on a Computer

In this section the various methods which can be used to model the dynamical evolution of stellar clusters, where one calculates how the distribution of stars within the cluster evolves over



■ Fig. 17-4

The various methods which can be employed to model the dynamical evolution of a stellar cluster (Fig. 3 from Amaro-Seoane et al. (2007), reproduced with permission)

time, are discussed. Four different methods are considered: (1) direct N-body calculations, (2) Fokker–Planck codes, (3) gas models, and (4) Monte Carlo codes. These approaches vary in both physical realism and speed, as shown in Fig. 17-4. In direct N-body calculations, stars are treated as individual objects and forces between all the stars are calculated directly. By definition, this is the most realistic approach but also the most expensive in terms of computer time taken to simulate the dynamical evolution of a cluster, with simulations taking several weeks on custom-built hardware designed for the purpose. In Fokker–Planck codes the evolution of a discretized version of the distribution function is followed numerically. This has the enormous advantage of speed (one can model the evolution of a cluster in typically less than 1 day), but this comes at the cost in realism concerning certain processes occurring in the cluster, such as stellar collisions and binary evolution. In the so-called gas models the cluster is modeled as a self-gravitating ball of (thermally conducting) gas, rather analogous to modeling a star. In a sense one is then considering a natural limit as the number of objects (stars) within the cluster becomes very large. Monte Carlo codes sit somewhere between the gas and Fokker–Planck methods on one side and the full N-body codes on the other. Stars sharing the same orbital and stellar properties are grouped together and treated as a single particle. The orbits of stars are then evolved over time via interactions between particles. This method is somewhat faster than a full N-body code (though slower than Fokker–Planck and gas methods) while still allowing for the implementation of processes such as stellar collisions. Each of the four methods are described in more detail below.

2.4.1 Full N-Body Codes

Why is it so difficult to model a stellar cluster completely, simply computing the gravitational forces between the individual stars and then accelerating the stars? As will be seen in the next section, two-body scattering is what drives dynamical evolution in clusters via dynamical relaxation. Thus any code must be able to include accurately the effects of many small perturbations (distant fly-bys by passing stars). One has to therefore be careful when using programmes which calculate approximations of the gravitational force (such as treecodes) and complete force calculations will be better (if more expensive). Additionally there is a huge range of timescales one has to consider when modeling the dynamical evolution of a stellar cluster. For a typical globular cluster, the size scale is a few pc, which implies crossing times of about 10^5 – 10^6 year. This should be compared to typical cluster ages of around 10^{10} year. But the range of scales is even larger: some of the objects within a cluster are in reality very tight binary star systems, having orbital periods of only hours. Evolving a stellar cluster therefore involves taking a great number of numerical integration steps. Over the years a great deal of development of N-body codes has taken place, most notably led by Sverre Aarseth (1999, 2003). Various techniques have been employed to make the problem more tractable: for example, treating very tight binaries in specific ways and using custom computer hardware such as GRAPE boards (e.g., Makino 1996) and computer graphics cards (GPUs). Currently, one is able to simulate a cluster of $\sim 3 \times 10^4$ stars completely; for example, the rich open cluster M67 has been modeled (Hurley et al. 2005).

2.4.2 Fokker–Planck Approach

In Fokker–Planck codes, one considers the evolution of the distribution function due to effects of perturbations such as two-body scattering. For stars moving in the smooth potential ϕ of the cluster (calculated via Poisson’s equation – see (► 17.3)), the evolution of the distribution function $f(\mathbf{r}, \mathbf{v}, t)$ is given by the collisionless Boltzmann equation (Binney and Tremaine 2008):

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{r}} - \frac{\partial \phi}{\partial \mathbf{r}} \cdot \frac{\partial f}{\partial \mathbf{v}} = 0 \quad (17.30)$$

where df/dt is the Lagrangian or full time derivative. Thus if the potential were perfectly smooth, f for a particular group of stars would not change as they orbited within the cluster. In a real stellar cluster, the potential is not smooth on all scales. It is this graininess in the potential that leads to relaxation, in other words they cause the distribution function f to evolve in time. These relaxational effects are also termed collisional effects (collisional here means that velocities are changing rather than meaning a physical impact), so one talks about an additional collisional term $\Gamma(f)$ such that $df/dt = \Gamma(f)$.

It is assumed that two-body scattering from close encounters is the cause of the relaxation. This occurs locally in the sense that a flyby changes the velocity of a star but not its position. It turns out then that (Binney and Tremaine 2008)

$$\Gamma(f) = - \sum_{i=1}^3 \frac{\partial}{\partial v_i} (D[\Delta v_i]f(\mathbf{r}, \mathbf{v}, t)) + \frac{1}{2} \sum_{i,j=1}^3 \frac{\partial^2}{\partial v_i \partial v_j} (D[\Delta v_i \Delta v_j]f(\mathbf{r}, \mathbf{v}, t)) \quad (17.31)$$

where $D[\Delta v_i]$ is the average change per unit time in v_i due to encounters. The evolution of the distribution function can be expressed and solved numerically using finite difference methods (e.g., see Cohn 1979; Chernoff and Weinberg 1990). Fokker–Planck codes have been used to model a number of observed clusters, for example NGC 6397 (Drukier 1995).

2.4.3 Gas Models

A cluster of stars held together by their mutual gravitational forces can behave in similar ways to a self-gravitating ball of gas, providing the number of stars is sufficiently large (e.g., Larson 1970a, b; Lynden-Bell and Eggleton 1980). It will be seen in [Sect. 3](#) how consideration of the thermodynamic properties of such an object will lead us to understand how the central cluster density evolves over time. For a cluster which is spherically symmetric, the structure of the cluster can be described by a density $\rho(r)$ and temperature $T(r)$ where temperature in effect replaces the velocity dispersion of the cluster σ , and the stellar cluster then has a pressure $p(r) = \rho\sigma^2$. One can then rewrite the equations of stellar structure but applied to a stellar cluster. These are (Lynden-Bell and Eggleton 1980; Heggie and Hut 2003)

$$\begin{aligned}\frac{\partial M}{\partial r} &= 4\pi\rho r^2 \\ \frac{\partial p}{\partial r} &= -\frac{GM(r)}{r^2}\rho \\ \frac{\partial L}{\partial r} &= -4\pi\rho r^2\left(\sigma^2\frac{dS}{dt} - \epsilon\right) \\ \frac{\partial\sigma^2}{\partial r} &= -\frac{1}{3GmC\ln\Lambda}\frac{\sigma}{\rho}\frac{L(r)}{4\pi r^2}\end{aligned}\tag{17.32}$$

where σ is the root mean square one-dimensional speed, m is the mass of an individual star, C is a constant connected with the relaxation process, $\ln\Lambda$ is known as the Coulomb logarithm (and is connected with two-body scattering as will be seen in [Sect. 3](#)), $S = \ln(\sigma^3/\rho)$, and ϵ is the rate of energy generation in three-body interactions (where an interaction between three unbound stars produces a binary and a third star with the excess energy). With the addition of thermal conduction, where heat flows from the cluster center to the outer regions, one may then model the dynamical evolution of the cluster. This works in part because the timescale for this process is rather long, as will be seen in [Sect. 3](#).

2.4.4 Monte Carlo Codes

The Monte Carlo numerical scheme is related to the Fokker–Planck method described earlier. However, rather than solve the Boltzmann equation (with the collision term) by finite difference methods, a particle approach is followed where the stars on similar orbits are grouped together and treated as a single particle. This approach was first introduced by Hénon (1971) in order to model the dynamical evolution of globular clusters. The method assumes the cluster is spherically symmetric and in dynamical equilibrium. In other words, one is able to take time steps which are much larger than orbital timescales instead using time steps which are some fraction of a relaxation time. The position of each particle along its orbit is chosen randomly, weighting

each position in the orbit with the time it spends there. One is free to have different time steps in different places within the cluster, for example, taking shorter time steps in the center of the cluster where the relaxation timescale is typically much shorter.

Relaxation is treated as a diffusive process in the same way as with the Fokker–Planck approach described earlier. As before, two-body scatterings give rise to a change of the orbits of the stars, or in the case here the particles representing a group of stars. At each step, a pair of particles is selected randomly, and their orbits are changed due to their mutual encounter. Over time, therefore, stars' orbits are changed by the effect of a large number of two-body encounters, as will be the case in the real cluster. An important advantage of the Monte Carlo approach compared to the Fokker–Planck method is that other processes can be included, for example, encounters involving binaries (Fregeau and Rasio 2007; Fregeau et al. 2003). Such encounters will be important for the dynamical evolution of clusters as will be seen in [Sect. 4](#). Monte Carlo codes have been used to model observed clusters, including M4 (Heggie and Giersz 2008) and NGC 6397 (Giersz and Heggie 2009).

3 Internal Dynamical Evolution

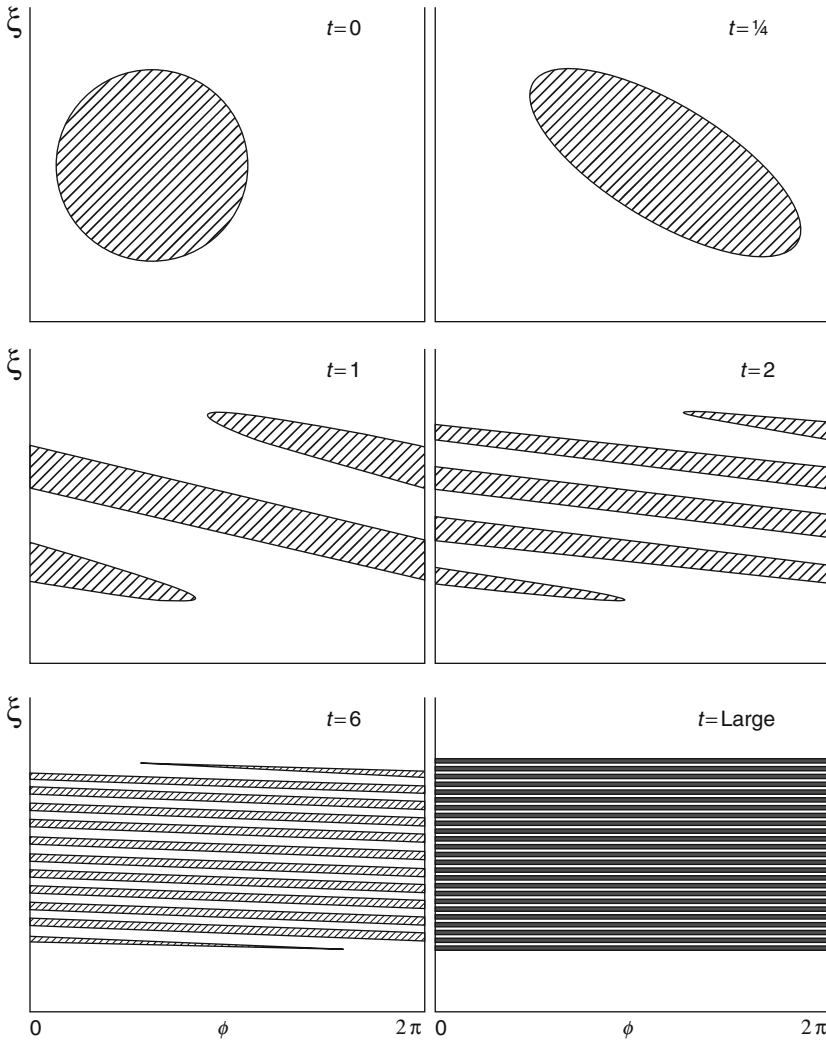
In this section the dynamical evolution of a stellar cluster is considered, including a description of the processes which occur within a cluster. The effects of the finite size of stars (which lead to physical collisions), the effects of primordial binaries, or the effects of a population of stellar-mass black holes are all neglected in this section. These effects will be dealt with in [Sect. 4](#). The effects of the galaxy on a cluster, such as tidal stripping and heating from disk and bulge shocks, are considered in the final section of this chapter.

3.1 Processes Occurring at the Early Stages

In this subsection the processes occurring at the very early stages in the cluster evolution are first considered. In the next subsection, two-body relaxation is discussed, where the accumulated effect of many scattering events between stars leads to energy transfer. It will turn out that this process occurs on a timescale of $10^8 - 10^{10}$ years for typical conditions (number densities and velocity dispersions) within a cluster. As will be seen, two-body relaxation is the driving force in the long-term cluster evolution. However there are other processes which have an impact very early on in the history of a cluster: (1) phase mixing, (2) violent relaxation, and (3) mass loss from the cluster, driven by stellar evolution. These three processes are discussed in more detail below.

3.1.1 Phase Mixing and Violent Relaxation

Consider some initial conditions where all the stars occupy a particular patch of phase space, while most of the phase space is empty. As the system evolves following the collisionless Boltzmann equation ([Eq. 17.30](#)), this patch of phase space is stretched, smeared and distorted. It will mix with the parts of the phase space which were initially empty. A sketch of this process is shown in [Fig. 17-5](#). So even though the microscopically small region of phase space around



■ Fig. 17-5

Figure showing the process of phase mixing where a small patch of phase space initially containing all the stars is stretched and distorted as the system evolves given by the collisionless Boltzmann equation and mixes with the region of phase space that does not contain stars (Fig. 2 from Lynden-Bell (1967), reproduced with permission)

a given star will conserve its original number density of stars, the larger region around it will experience a reduction in the number density of stars as the initial number density is averaged together with regions containing no stars (Lynden-Bell 1967).

Phase mixing does not change the energy of a particular star as it orbits within the cluster. However, in a process known as violent relaxation, the energy of particular stars is changed if the potential is a function of time as well as position (Lynden-Bell 1967). This will be the case,

for example, when a cluster is forming out of various lumps of gas which are falling together and merging in a messy and complicated way. The rate of change of energy is then given by (Binney and Tremaine 2008)

$$\frac{dE}{dt} = \left(\frac{\partial \Phi}{\partial t} \right)_{\text{trajectory}} \quad (17.33)$$

where the right-hand term is the rate of change of potential with respect to time along the trajectory followed by a particular star. One can see how a changing potential can lead to a change in the total energy of a particular star by considering the following examples (Binney and Tremaine 2008). Firstly, imagine a star located in the center of a cluster where material is falling in to cluster from outside, being fed directly into the central regions. The deeper potential well produced from the mass infall will mean that the star is now more deeply bound, i.e., its total energy has decreased. In other cases, stars can have their energies increased, for example, a star plunging through the middle of the cluster from the halo may be accelerated inward by a concentrated mass distribution but decelerated by a less centrally concentrated mass distribution (i.e., receive less net deceleration) if the object expands while the star passes through the central regions. Importantly, the process of violent relaxation is independent of the mass of a particular star, in other words the velocity distribution is independent of stellar mass, i.e., $v^2 = \text{a constant}$ (Lynden-Bell 1967). This is different from the equipartition which follows from two-body relaxation where $mv^2 = \text{a constant}$, where m is the mass of a particular star.


3.1.2 Mass Loss from the Cluster

The effects of mass loss within a cluster are now considered. When stars form in a cluster from a cloud of gas, the star formation process is not completely efficient and some fraction of the gas is not converted into stars but remains as gas within the cluster. This gas might be stripped from the globular cluster when it passes through the galactic disk. Alternatively the gas may be ejected from cluster, driven out by massive stars within the cluster, either via their winds or supernova explosions. This will occur on a timescale of about 10^7 years, which is much shorter than the orbital timescale of the cluster within the galaxy. Thus the evolution of massive stars is more likely to lead to gas ejection from the cluster. If the star formation efficiency has been rather low when a particular cluster formed, the rapid ejection of the remaining gas, driven out, for example, by supernova explosions, could unbind the cluster or at least lead to ejection of a significant fraction of the stars as much of the gravitational binding mass is lost (Baumgardt and Kroupa 2007; Kroupa et al. 2001).

Mass loss from the stars themselves due to stellar evolution can also effect the cluster. As stars evolve off the main sequence they expand to become red giants. These red giants later evolve to eject their envelopes, either violently as supernova explosions in the case of more massive stars ($m \geq 8 M_{\odot}$) producing neutron stars or stellar-mass black holes, or more steadily for other stars leaving white dwarfs as remnants. The slow ejection of mass in stellar winds is different from the rapid ejection in a supernova explosion. To simplify the discussion here, consider the case of Jupiter as it orbits around the Sun. If the sun were to lose more than half its mass in an instant, Jupiter would fly off unbound as its orbital speed would exceed the escape speed around the reduced Sun. However, if the same amount of mass loss were to occur on *long timescales* then in fact Jupiter's orbit would simply expand by a factor of two but maintain its original eccentricity, as the product of the orbit's semi-major axis and the total mass in the

system is conserved. Slow mass loss will cause the cluster to expand rather than fly apart. Even a relatively small fraction of mass being lost in this way can cause a significant expansion of the cluster (Binney and Tremaine 2008). Fokker–Planck modeling of clusters shows that mass loss due to stellar evolution in the first 5×10^9 years could lead to the disruption of some (weakly bound) clusters when one considers the galactic tidal field, with the slope of the initial mass function having a significant effect (Chernoff and Weinberg 1990). Initial mass segregation also plays an important role, with clusters which are initially more mass segregated (i.e., those having a larger fraction of the most massive stars located in the cluster center) suffering from a greater expansion when stars lose mass through evolution (Vesperini et al. 2009).

3.2 Two-Body Relaxation

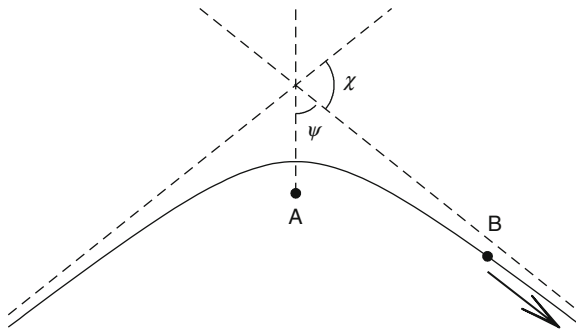
Consider first a single scattering event, as shown in  Fig. 17-6, where the trajectory of a passing star is plotted in the frame of the star it encounters.


Let us consider two stars A and B of mass m_1 and m_2 , encountering each other with an impact parameter p and a relative speed at infinity v_∞ . The angular deflection of the orbit is given by χ . It turns out that it is easier to consider the angle ψ as shown in the figure, which is related to χ via the equation $\chi = \pi - 2\psi$. It can be shown (Spitzer 1987) that

$$\tan \psi = \frac{pv_\infty^2}{G(m_1 + m_2)} \equiv \frac{p}{p_0} \quad (17.34)$$

where p_0 is the impact parameter for encounters for which the deflection $\chi = \pi/2$. If the two stars are moving at a relative speed at infinity v_∞ , the shift in the velocity of star 1 is given by

$$\begin{aligned} \Delta v &= 2v_\infty \sin(\chi/2) \cdot \frac{m_1}{m_1 + m_2} = 2v_\infty \sin(\pi/2 - \psi) \cdot \frac{m_1}{m_1 + m_2} \\ &= 2v_\infty \cos\psi \cdot \frac{m_1}{m_1 + m_2} \end{aligned} \quad (17.35)$$



 Fig. 17-6

A figure showing a fly by encounter between two stars A and B, considered in the rest frame of A

which using (17.34) can be rewritten as

$$(\Delta v)^2 = \frac{4m_1^2 v_\infty^2}{(m_1 + m_2)^2} \frac{1}{1 + (p/p_0)^2} \quad (17.36)$$

To calculate $\langle (\Delta v)^2 \rangle$ one must integrate (17.36) for a range of impact parameters from head-on encounters to some maximum impact parameter p_{\max}

$$\begin{aligned} \langle (\Delta v)^2 \rangle &= \int_0^{p_{\max}} (\Delta v)^2 2\pi p n v_\infty dp \\ &= \frac{4\pi G^2 n m^2}{v_\infty} \ln[1 + (p_{\max}/p_0)^2] \end{aligned} \quad (17.37)$$

Close encounters (those for which $p < p_0$) will lead to large deflections, while distant encounters (i.e., those for $p > p_0$) will lead to much smaller deflections. However the latter will be much more frequent. The combined effect of the more frequent distant encounters is in fact larger than the effect of the rarer close encounters. This means that the choice of p_{\max} is very important, $\Lambda = \ln(p_{\max}/p_0)$ being known as the Coulomb logarithm. For typical globular clusters $m \simeq 0.7 M_\odot$ and $v_\infty \simeq 8$ km/s, which imply $p_0 \simeq 9 \times 10^{-5}$ pc. Assuming $p_{\max} = r_c = 0.3$ pc, this implies $\Lambda = \ln(p_{\max}/p_0) = 8$. Hence one may make the approximation that

$$\ln[1 + (p_{\max}/p_0)^2] \simeq 2\ln\Lambda \quad (17.38)$$

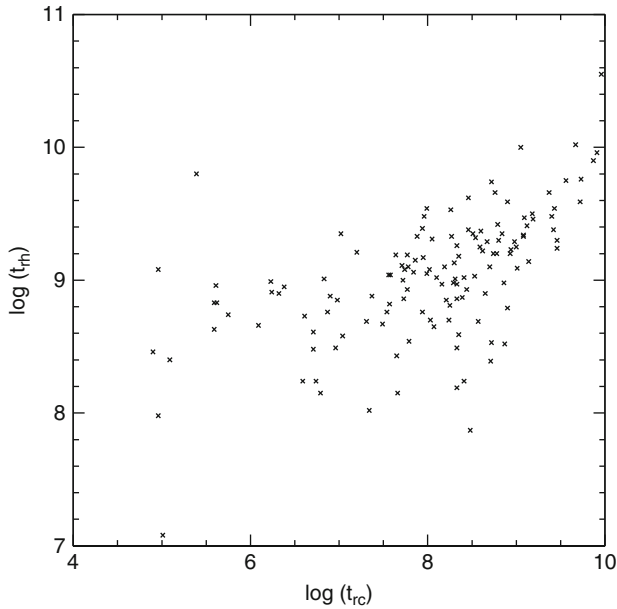
The two-body relaxation timescale is the time required for the accumulated deflections due to two-body scatterings to be significant (i.e., comparable) to the speed of the stars. The two-body relaxation timescale can be written as (Binney and Tremaine 2008)

$$\begin{aligned} t_{\text{relax}} &= \frac{1}{3} \frac{v_\infty^2}{\langle (\Delta v)^2 \rangle} = \frac{1}{24\pi G^2} \frac{v_\infty^3}{\ln\Lambda} \frac{1}{n m^2} \\ &= \frac{1.8 \times 10^{10} \text{ year}}{\ln\Lambda} \left(\frac{\sigma}{10 \text{ km/s}} \right)^3 \left(\frac{M_\odot}{m} \right) \left(\frac{10^3 M_\odot/\text{pc}^3}{\rho} \right) \end{aligned} \quad (17.39)$$

It is important to understand that the relaxation time is a function of location within a stellar cluster (as both density and velocity dispersion are functions of radius within a cluster). One could, for example, compute the relaxation time at the cluster core by inserting the core properties into (17.39). Another convenient relaxation time for a cluster is the median relaxation time t_{rh} (also known as the half-mass relaxation time) which is obtained by replacing the density in (17.39) with the mean density inside the cluster's half-mass radius r_h and by replacing $3\sigma^2$ with the mean-square speed of the cluster's stars $\langle v^2 \rangle \simeq 0.4 GM/r_h$ (Binney and Tremaine 2008). One can also approximate $p_{\max} \sim r_h$ and p_0 from (17.34), thus $\Lambda = r_h \langle v^2 \rangle / Gm = 0.4N$. One therefore obtains

$$\begin{aligned} t_{\text{rh}} &= \frac{0.14N}{\ln(0.4N)} \sqrt{\frac{r_h^3}{GM}} \\ &= \frac{6.5 \times 10^8 \text{ year}}{\ln(0.4N)} \left(\frac{M}{10^5 M_\odot} \right)^{\frac{1}{2}} \left(\frac{1 M_\odot}{m} \right) \left(\frac{r_h}{1 \text{ pc}} \right)^{\frac{3}{2}} \end{aligned} \quad (17.40)$$

In Fig. 17-7 the half-mass relaxation time t_{rh} is plotted as function of the core relaxation time t_{rc} for all observed globular clusters in our galaxy, with the cluster data being taken



■ Fig. 17-7

Plot showing the log of the core relaxation timescale t_{rc} as a function of the log of the half-mass relaxation timescale t_{rh} (both in years) for globular clusters with data taken from the Harris Catalogue

from the Harris Globular Cluster Database (which is available online at <http://www.physics.mcmaster.ca/Globular.html>). For virtually all clusters $10^8 \leq t_{rh} \leq 10^{10}$ years, whereas the core relaxation time t_{rc} can be much shorter in some cases. One can see from [Fig. 17-7](#) that there is a correlation between t_{rc} and t_{rh} : clusters with shorter half-mass relaxation times tend to have much shorter core relaxation times. More will be seen in detail why this is in [Sect. 3.4](#). The key point is that two-body relaxation drives the dynamical evolution of the cluster: on a timescale of $\sim 10\text{--}20 t_{rh}$, the cluster core contracts to reach very high densities. As can be seen from [17.39](#), the core relaxation time will go down as the central density increases. Thus the clusters having shorter core relaxation times are generally those that are more centrally concentrated.

3.3 Mass Segregation

Energy will be exchanged between stars via two-body scattering. Such encounters will leave the various stellar populations (i.e., stellar masses) having similar kinetic energies at the same location within the cluster. Therefore the stars with more mass will tend to sink toward the cluster core, while the lighter stars are more likely to be found further out in the cluster or to escape. This process of mass segregation will take place on the relaxation timescale. As can be seen from [Fig. 17-7](#), where the half-mass and core relaxation timescales are plotted for all observed globular clusters in our galaxy, one would expect some clusters to be more segregated than others.

Let us suppose a cluster contains two stellar species, having individual masses m_1 and m_2 , where $m_2 > m_1$, and the total mass contained in each population is M_1 and M_2 . In order that the two populations can achieve equipartition one requires (Spitzer 1969, 1987)

$$\frac{M_2}{M_1} \left(\frac{m_2}{m_1} \right)^{3/2} < 0.16 \quad (17.41)$$

If the population of the more massive stars is too large, then they will sink into the middle of the cluster and form their own, separate cluster in the middle of a core of the lower-mass stars. This process is known as the equipartition instability. One particular example where this instability occurs is the case where a cluster retains a significant population of stellar-mass black holes (e.g., Mackey et al. 2008). Such objects are considerably more massive than white dwarfs, neutron stars, or any main-sequence stars found in old globular clusters today. The evolution of clusters containing a separate, central population of stellar-mass black holes will be discussed in [Sect. 4](#).

3.3.1 Observations of Mass Segregation

There have been a great many observational studies of mass segregation within globular clusters. For a detailed discussion of this topic, the reader is directed to the chapter on the contents of globular clusters by Giampaolo Piotto.

In brief here, it will be noted that data taken with the Hubble Space Telescope has been instrumental in the study of mass segregation within globular clusters. The improved resolution (compared to ground-based telescopes) means that individual stars can be resolved even in the cores of many globular clusters. Photometric studies of clusters then yield information about the stellar population and how it varies as a function of radius within a cluster. Thus the effects of mass segregation can be measured by observing, for example, an excess of turn-off mass stars and a depletion of low-mass stars in the cluster cores. In this way, observational evidence of mass segregation has been seen in a number of globular clusters, including: NGC 6752 (Shara et al. 1995), 47 Tucanae (Paresce et al. 1995a), and NGC 6397 (King et al. 1995, 1996).

Observations of the trapezium cluster in Orion show that most of the most massive stars are already located in the center of the cluster despite the relative dynamical youth of the system. N-body calculations suggest that the cluster had primordial mass segregation, i.e., that the massive stars were formed in the middle (Bonnell and Davies 1998). Such primordial mass segregation would have an important effect on the subsequent evolution of a globular cluster (Mackey et al. 2008, and [Fig. 17-11](#)).

3.4 Core Evolution

3.4.1 The Gravothermal Catastrophe

A self-gravitating system has a negative heat capacity. In other words when the system loses energy, it gets hotter in the sense that the stars' velocities increase. This can be seen easily via

the virial theorem (e.g., Binney and Tremaine 2008). The temperature T of a self-gravitating system can be defined via the equation

$$\frac{1}{2}m\overline{v^2} = \frac{3}{2}k_B T \quad (17.42)$$

where m is the stellar mass and k_B is Boltzmann's constant. The total kinetic energy of a system of N stars is

$$K = \frac{3}{2}Nk_B\overline{T} \quad (17.43)$$

where \overline{T} is the mass-weighted mean temperature. From the virial theorem one can know the total energy of the system $E = -K$. Hence one has

$$E = -\frac{3}{2}Nk_B\overline{T} \quad (17.44)$$

The heat capacity of the system is then simply

$$\frac{dE}{d\overline{T}} = -\frac{3}{2}Nk_B \quad (17.45)$$

which indeed is negative. What does this imply for the possible evolution of a self-gravitating cluster of stars? In a simple picture consider the cluster as being made up of two components: a central core surrounded by a halo. If the core becomes hotter than the halo, heat (i.e., energy) will flow from the core to the halo. In practice this heat flow will take place via two-body scattering. Because the core has a negative heat capacity, as it shrinks due to energy loss its temperature will increase. This increase in temperature will lead in turn to an increase in the energy loss rate. Thus the process will continue to accelerate, with the core contracting to reach very high densities. This instability is known as the gravothermal catastrophe. In practice the core and halo are not separate components. However, detailed analyses show that this instability can occur in real systems (Lynden-Bell and Wood 1968; Spitzer 1969).

3.4.2 Core Collapse

If one models the evolution of an isolated stellar cluster one typically finds that the outer regions of the cluster expand, while the core contracts (by the processes of mass segregation and gravothermal instability), reaching extremely high densities on a timescale around $10\text{--}20 t_{\text{rh}}$ (where t_{rh} is the half-mass relaxation timescale as given in (17.40)). This process is known as core collapse. Relating the central density ρ_0 to the core radius r_0 , as a function of time, one finds that $\rho_0(t) \propto r_0^{-\alpha}(t)$. Cohn (1980) found $\alpha = 2.23$ using a Fokker–Planck code, while Lynden-Bell and Eggleton (1980) found $\alpha = 2.21$ using a gas model for a stellar cluster.

Following the approach given in Binney and Tremaine (2008), one is able to produce scaling relations for the core properties as the cluster approaches the time of core collapse. If one defines the core mass $M_0(t) = \rho_0(t)r_0^3(t)$, then $M_0 \propto r_0^{3-\alpha}$. Choosing r_0 to be the King radius, i.e., that $r_0 = (9\sigma^2/4\pi G\rho_0)^{1/2}$, then the central velocity dispersion is $\sigma \propto \rho_0^{1/2}r_0 \propto r_0^{\alpha/2-3}$. From (17.39) the relaxation time in the core is $t_{\text{relax}} \propto \sigma^3/\rho_0 \propto r_0^{3-\alpha/2}$ where $\ln\Lambda$ has been taken to be a constant, which is not a bad approximation.

The process of core collapse is driven by relaxation and therefore one can reasonably assume that the timescale for changes in the core radius should be the same as the relaxation timescale. In other words

$$\frac{1}{r_0} \frac{dr_0}{dt} \propto \frac{1}{t_{\text{relax}}} \propto r_0^{\alpha/2-3} \quad (17.46)$$

This equation can be solved giving

$$r_0(t) \propto (t_{\text{cc}} - t)^{2/(6-\alpha)} \propto \tau^{0.53} \quad (17.47)$$

where $\alpha = 2.23$; t_{cc} is the moment of core collapse and $\tau = t_{\text{cc}} - t$ is the time remaining until core collapse. One also has (Binney and Tremaine 2008)

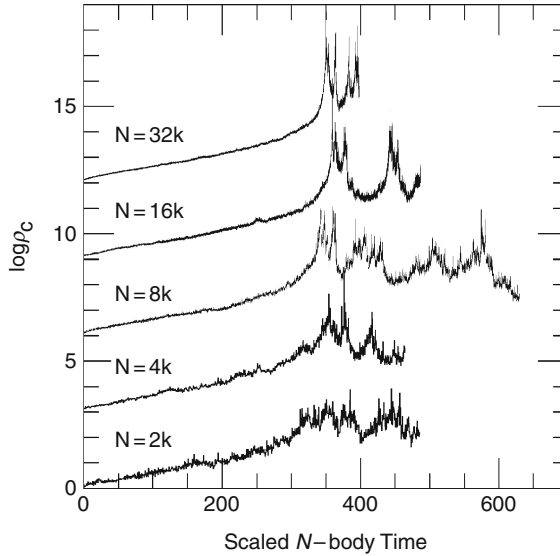
$$\begin{aligned} \rho_0(t) &\propto \tau^{-2\alpha/(6-\alpha)} \propto \tau^{-1.18} \\ \sigma^2(t) &\propto \tau^{(4-2\alpha)/(6-\alpha)} \propto \tau^{-0.12} \\ M_0(t) &\propto \tau^{(6-2\alpha)/(6-\alpha)} \propto \tau^{0.41} \\ t_{\text{relax}}(r=0) &\propto \tau \end{aligned} \quad (17.48)$$

It is found that the time to core collapse $\tau \simeq 300t_{\text{rc}}$ (or equivalently about 10–20 t_{rh}) for clusters containing stars of a single mass (see Table 1 of Quinlan 1996). For clusters containing stars of unequal mass, the time to core collapse is reduced (Chernoff and Weinberg 1990). One should also note from (● 17.48) that although the density increases steeply, the mass contained in the core M_0 in fact is decreasing as the moment of core collapse is approached.

Core collapse can be halted provided there is a source of energy injection within the core to offset the loss of energy from the core into the cluster halo. It turns out that scatterings involving binaries form such an energy source. When a tight binary star system is involved in a flyby encounter with a single star, the binary is hardened (becomes more bound) while the single star receives a kick that increases its kinetic energy (as total energy is conserved). This process provides an input of energy to the core which can arrest the collapse. If the cluster does not originally contain any binaries, then a population will be produced close to core collapse by three-body interactions where three unbound single stars interact, leaving two in a binary with the third star carrying off the excess energy. In this case energy is also injected into the core when the binaries are formed. In addition to binaries formed via three-body encounters, the cluster may contain primordial binaries, or binaries formed via tidal interactions. In both cases, these binary populations will also inject energy into the core and thus halt core collapse, or even in some cases prevent its onset. This will be discussed in more detail in ● Sect. 4.

3.4.3 Gravothermal Oscillations

The evolution of the cluster after core collapse is now considered. Binaries in the core give kicks to passing single stars via binary–single scattering. These stars then spread this extra energy throughout the cluster via two-body scattering. The cluster expands self-similarly with a scale-length $R \propto \tilde{\tau}^{2/3}$ where $\tilde{\tau}$ is the time since core collapse (Goodman 1984). As the cluster expands, the central density decreases, and the rate of energy injection into the cluster from the binaries in the core also drops. It turns out that the cluster core regulates itself such that self-similar expansion is maintained.



■ Fig. 17-8

Logarithm of the cluster central density as a function of time for N -body simulations of clusters containing various numbers of stars. Gravothermal oscillations (spikes in the density at 350 time units and later) are clearly visible for clusters containing more than 8,000 stars (Fig. 1 from Makino (1996), reproduced with permission)

This expansion of the cluster, driven by binary heating in the core, is unstable for clusters containing a sufficient number of stars. In this case if the core expands, and cools, it turns out that material just outside the core is now hotter than the core. Heat now flows from these surrounding regions into the core, driving further core expansion and cooling. In the end, the core comes into contact with stars much further out in the cluster which have a lower temperature than the core stars. The core can now transfer energy outward again. By now the core has expanded considerably and must undergo another cycle of gravothermal collapse in order to reach higher densities and so generate the energy input required by the cluster. These cycles are known as gravothermal oscillations. They have been seen in Fokker-Planck simulations (Gao et al. 1991), N -body simulations (Makino 1996), and Monte Carlo simulations (Fregeau et al. 2003; Heggie and Giersz 2008). In [Fig. 17-8](#) the logarithm of the central density is shown as a function of time for the simulations of Makino et al. Gravothermal oscillations become apparent for simulations involving a larger number of stars.

4 Complications and Additional Effects

In this section the various additional effects are considered which will act as complications to the internal evolutionary processes described in [Sect. 3](#). The frequency and role of physical collisions between stars is discussed. The production of binaries via tidal interactions between stars is considered. Tidal-capture and primordial binaries provide an important heat source

for globular clusters via binary–single scattering. Dynamical interactions involving binaries can also be an important formation channel for exotic objects such as neutron star binaries. Stellar-mass black holes, should they be produced and retained in globular clusters, could cause cluster expansion, and might produce intermediate-mass black holes through runaway mergers. Observational evidence suggests that many globular clusters contain multiple stellar populations. If true, this would represent a paradigm shift in our picture of globular cluster formation.

4.1 Stellar Collisions

Close encounters between two stars, where the two stars pass within a few stellar radii of each other, are extremely rare in the low-density environment of the solar neighborhood. However, in the cores of globular clusters, and galactic nuclei, number densities are sufficiently high ($\sim 10^5$ stars/pc³ in some clusters) that encounter timescales can be comparable to, or even less than, the age of the universe (Hills and Day 1976). In other words, a large fraction of the stars in these systems will have suffered from at least one close encounter or physical collision in their lifetime.

The cross section for two stars, having a relative velocity at infinity of v_∞ , to pass within a distance R_{\min} is given by

$$\sigma = \pi R_{\min}^2 \left(1 + \frac{v^2}{v_\infty^2} \right) \quad (17.49)$$

where v is the relative velocity of the two stars at closest approach in a parabolic encounter (i.e., $v^2 = 2G(m_1 + m_2)/R_{\min}$, where m_1 and m_2 are the masses of the two stars). The second term is due to the attractive gravitational force between the two stars, and is referred to as gravitational focusing. In the regime where $v \ll v_\infty$ (as might be the case in galactic nuclei with extremely high velocity dispersions), one recovers the result, $\sigma \propto R_{\min}^2$. However, if $v \gg v_\infty$ as will be the case in systems with low velocity dispersions, such as globular clusters, $\sigma \propto R_{\min}$.

One may estimate the timescale for a given star to undergo an encounter with another star, $\tau_{\text{coll}} = 1/n\sigma v$. For clusters with low velocity dispersions, one thus obtains

$$\tau_{\text{coll}} = 7 \times 10^{10} \text{ year} \left(\frac{10^5 \text{ pc}^{-3}}{n} \right) \left(\frac{v_\infty}{10 \text{ km/s}} \right) \left(\frac{R_\odot}{R_{\min}} \right) \left(\frac{M_\odot}{m} \right) \text{ for } v \gg v_\infty \quad (17.50)$$

where n is the number density of single stars of mass m . For an encounter between two single stars to be hydrodynamically interesting, one typically requires $R_{\min} \sim 3R_\star$ for $v_\infty = 10$ km/s (see for example, McMillan et al. 1987). It is thus seen that for typical globular clusters, where $n \sim 10^5$ stars/pc⁻³, up to 30% of the stars in the cluster cores will have undergone a collision or close encounter at some point during the lifetime of the cluster.

Collisions and close encounters will be important for a number of reasons. They may produce the various stellar exotica which have been observed in clusters, such as blue stragglers and millisecond pulsars (e.g., Bailyn 1995). Stellar collisions will also play a role in the dynamical evolution of clusters as mass loss due to stellar evolution will be enhanced (owing to prompt mass loss during collisions and also the shorter lifetimes of the more massive stars produced via the merger of two lower-mass stars).

Blue stragglers appear to be main-sequence stars that are more massive than the current turn-off mass in clusters. They have been observed in all clusters (e.g., Piotto et al. 2004). One

natural explanation is that they are formed via the merger of two lower-mass main-sequence stars in a collision. Hydrodynamic simulations of such collisions, and modeling of the subsequent evolution, suggest that this is a viable formation channel (Benz and Hills 1987; Sills et al. 1997). However the number of blue stragglers seen within clusters seems not to scale simply with the expected collision rate, indicating that this is not the only formation channel. Alternatively, blue stragglers may be formed via mass transfer within binaries where mass from an evolving primary increases the mass of the secondary. A mixture of these two channels seems able to explain the observed population (Davies et al. 2004). Both channels have been seen to occur in N-body models of M67 (Hurley et al. 2001, 2005).

Collisions between red giants and compact objects (black holes, white dwarfs, or neutron stars) might be one way to produce compact, interacting binaries. A physical collision will leave a compact impactor orbiting the red-giant core, both engulfed in a common envelope formed from the red-giant envelope. The envelope will be ejected as the binary within tightens. Depending on the envelope mass, either the red-giant core and the impactor will spiral in and merge, or they will eject the entire envelope, leaving being a very tight binary containing the red-giant core (essentially a white dwarf) and the compact-object which struck the red giant (Davies et al. 1991; Rasio and Shapiro 1991).

4.2 Stellar Binaries

4.2.1 Making Binaries via Tidal Capture

It has been seen in the previous subsection how stellar collisions are relatively frequent in the dense cores of globular clusters. Close encounters also occur, where two stars pass within a few stellar radii of each other. During such an encounter, non-radial oscillations will be excited within at least one of the stars, the required energy being taken from the orbit. Given that the velocity dispersion in globular clusters is rather modest (around 10 km/s compared to the surface escape speed of the sun of some 618 km/s), close encounters can result in tidal capture to form a tight binary.

By considering the oblateness induced within a star by tides, Fabian et al. (1975) estimated that tidal capture of a compact object (neutron star, black hole, or white dwarf) would occur when it passed within approximately three stellar radii of a main-sequence or red-giant star for $v_\infty \sim 10$ km/s which is typical for globular clusters. More detailed calculations obtain similar cross sections for tidal capture to occur (Lee and Ostriker 1986; Press and Teukolsky 1977). Initially binaries produced by tidal capture will be rather eccentric. Circularization will produce systems with a semi-major axis approximately twice the initial percenter separation (i.e., up to six stellar radii). If the stars can survive both the capture and circularization phases without merging, then an interacting binary is likely to be produced where mass flows from the main-sequence or red-giant star onto the compact object, for example, forming a low-mass X-ray binary (LMXB) in the case where the compact object is a neutron star or black hole. The relatively high numbers of LMXBs seen in clusters may therefore have been produced by tidal captures (Verbunt and Hut 1987). Indeed the number of X-ray objects seen in a particular cluster scales with the encounter rate within that cluster (Pooley et al. 2003; see [Fig. 17-9](#)). However, it is unclear whether a binary formed by tidal capture can avoid merging as the energy injected into the stellar envelope is large. It is possible that rather than form a tight binary,

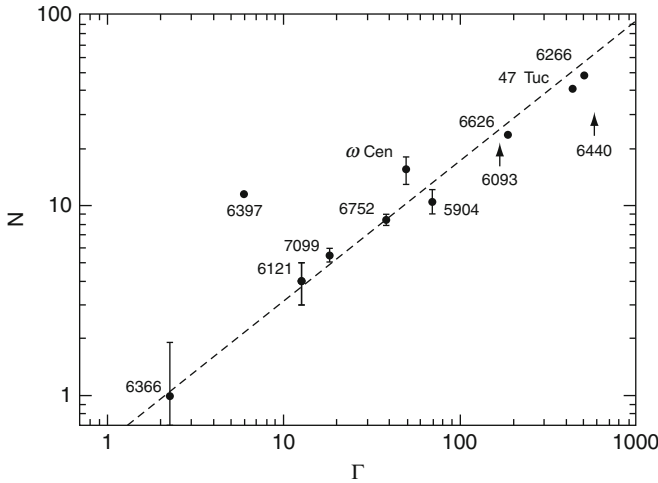


Fig. 17-9

Number of globular cluster X-ray sources (with $L_x \geq 4 \times 10^{30}$ ergs/s) as a function of the encounter rate between two single stars in the cluster (Fig. 2 from Pooley et al. (2003), reproduced with permission)

the non-compact star will simply expand and engulf the compact object forming a common-envelope system (McMillan et al. 1987). As will be seen in later subsections, an alternative formation channel is binary exchange involving single compact objects and primordial binaries containing two main-sequence stars, where an encounter results in the ejection of one of the main-sequence stars leaving the other main-sequence star in a relatively tight binary with the compact object. The scaling relationship between X-ray population and encounter rate seen by Pooley et al. would still apply as the encounter rate between binaries and single star scales with cluster properties in the same way as the encounter rate between two single stars.

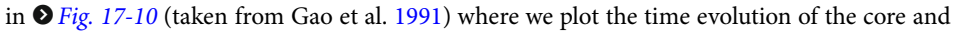
4.2.2 Binary/Single Encounter Dynamics and Heating

One may estimate the timescale for an encounter between a binary and a third, single star using (17.50) and taking $R_{\min} \approx a$, where a is the semi-major axis of the binary. The encounter timescale for a binary may therefore be relatively short as the semi-major axis can greatly exceed the size of the stars it contains. For example, a binary with $d \sim 1$ AU (i.e., $215 R_{\odot}$), will have an encounter timescale $\tau_{\text{enc}} \ll 10^{10}$ years in the core of a dense globular cluster. Thus encounters between binaries and single stars may be important in stellar clusters even if the binary fraction is small.

Encounters between single stars and extremely wide binaries will lead to the break up of the binaries as the kinetic energy of the incoming star exceeds the binding energy of the binary. Such binaries are often referred to as being *soft*. Conversely, *hard* binaries will be resilient to break up. They will in fact tend to be hardened further via encounters with other stars (Heggie 1975; Hills 1975). For typical cluster properties, the hard/soft boundary is between 3 and 10 AU.

But because of the progressive hardening of binaries through encounters with other stars, the widest binaries in cluster cores today will be around 1 AU only.

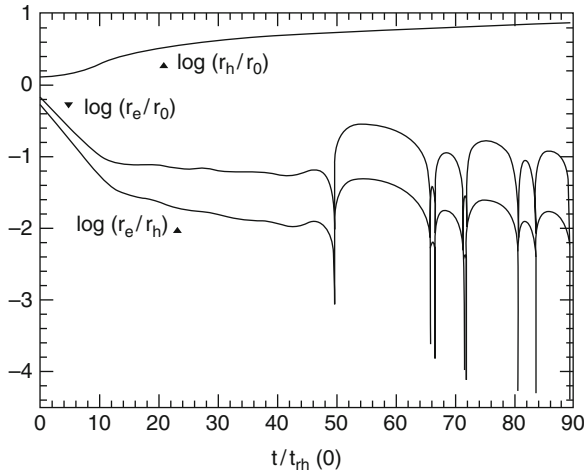
Encounters between single stars and hard binaries have three main outcomes (e.g., Hut and Bahcall 1983; Sigurdsson and Phinney 1995). A flyby may occur where the incoming third star leaves the original binary intact. However such encounters will harden (i.e., shrink) the binary and alter its eccentricity. Alternatively, an exchange may occur where the incoming star replaces one of the original components of the binary. During the encounter, two of the stars may pass so close to each other that they merge or form a very tight binary (as they raise tides in each other). This third channel is an important channel for the production of various varieties of stellar exotica.

As hard binaries are hardened through stellar encounters, they increase the kinetic energy of the stars they encounter. This energy will be redistributed to other stars via two-body scattering. Thus the cluster as a whole will be heated. The total binding energy contained within binary stars can exceed the binding energy of the stellar cluster. Therefore hardening of binaries represents an important energy source for clusters and may offset the moment of core collapse while the binaries present in the cluster are ground down (Gao et al. 1991). This is illustrated in  Fig. 17-10 (taken from Gao et al. 1991) where we plot the time evolution of the core and half-mass radii resulting from their Fokker–Planck calculations of a cluster whose initial conditions were based on a Plummer model. The ratio of the core radius to the half-mass radius, r_c/r_h , is kept relatively constant from 10 to about 50 half-mass relaxation times while the population of primordial binaries is consumed. Similar results are seen in N-body calculations (Heggie et al. 2006; Hurley 2007) and Monte Carlo calculations (Fregeau and Rasio 2007). However, all these calculations predict values of r_c/r_h around 0.03–0.05, whereas the observed distribution for clusters peaks around $r_c/r_h \sim 0.5$. It could be that most clusters are not yet in the binary-burning phase, or that additional energy sources, such as stellar or intermediate-mass black holes, may be heating the cluster cores (e.g., Mackey et al. 2007, 2008).

4.2.3 The Primordial Binary Population

The population of primordial binaries within a cluster plays an important role in the production of stellar exotica. Neutron stars and white dwarfs may exchange into primordial binaries, replacing main-sequence stars, forming X-ray binaries and cataclysmic variables (Verbunt and Hut 1987). Because binaries are on average heavier than the average star within a cluster, they will sink toward the cluster core due to mass segregation via the effects of two-body scattering. Within the dense core, they will interact with single stars, and other binaries. Wider (soft binaries) will be broken up and even hard binaries are vulnerable to break up in binary/binary encounters. Those binaries which are not broken up will be hardened and receive recoil kicks which will remove them (at least temporarily) from the cluster core. In some cases, recoil kicks may remove a binary from the cluster, although binaries containing non-compact stars are more likely to merge first.

The evolution of a primordial binary population within a cluster has been explored extensively using Monte Carlo techniques (e.g., Hut et al. 1992b; Ivanova et al. 2005) and via N-body simulations (Hurley et al. 2007). The binary population is seen to reduce over time as binaries are broken up or merge. Thus the binary fraction observed today will be lower than the initial value.



■ Fig. 17-10

The evolution of cluster core radius, r_c , and half-mass radius, r_h , as a function of time in units of the cluster's initial half-mass relaxation time. r_0 is the scale length for the initial Plummer model – labeled a in (17.8) (Fig. 1 from Gao et al. (1991), reproduced with permission)

Encounters in stellar clusters involving binaries produce a myriad of objects (Davies 1995; Davies and Benz 1995; Ivanova et al. 2006, 2008). In addition to exchange encounters directly producing low-mass X-ray binaries (LMXBs) and cataclysmic variables (CVs), encounters may lead to stellar mergers. These merger products may expand and engulf any companion star forming a common-envelope system, where the companion star and stellar core of the merger product will spiral together ejecting the envelope, leaving a much more compact binary. This may be a very important channel to produce tight binaries containing two compact objects. For example, globular clusters may be important sites for the production of tight binaries containing two neutron stars, which will merge via the effects of gravitational radiation. In some cases, an observed population can be a combination of the primordial population (i.e., the ones which would have also been seen outside of clusters) and one produced via dynamical encounters. One example of this is the cataclysmic variable population. In less-dense clusters, systems produced via the unmodified evolution of primordial binaries will dominate, while in the cores of the densest clusters, these relatively wide binaries would have been broken up via encounters with single stars, and any observed cataclysmic variables would have been formed either via tidal capture or from the exchange of single white dwarfs into harder primordial binaries (Davies 1997).

4.3 Black Holes

4.3.1 Stellar-Mass Black Holes

Black holes having masses around $5\text{--}10 M_{\odot}$ may be formed in core collapse supernovae of stars above $20\text{--}30 M_{\odot}$. Thus a massive globular cluster might contain several hundred stellar-mass

black holes, providing all those produced in core-collapse supernovae are retained. This is not a given: neutron stars are believed to receive natal kicks of typically several hundred km/s, which is far in excess of the escape speed from the center of a globular cluster (typically 40–60 km/s), and black holes may well receive kicks of similar magnitude, or at least similar momentum.

Binary companions may help retain both black holes and neutron stars in clusters. If the newly-formed black hole or neutron star remains bound to its stellar companion, the resulting speed of the binary will be much less than the kick given to the compact object, thus increasing the chances of keeping the compact object within the cluster. Massive stars tend to be in binaries with other massive stars, so this may be an effective way of retaining stellar-mass black holes.

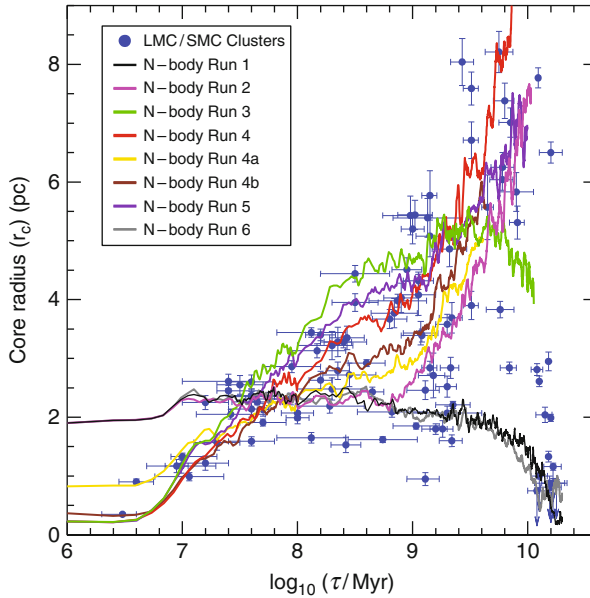
The effect a population of stellar-mass black holes is now considered in terms of the production of exotic objects and on the dynamical evolution of a stellar cluster. It turns out that black holes are relatively unlikely to exchange into binaries to form LMXBs in a manner similar to neutron stars. Indeed none of the dozen or so LMXBs observed in globular clusters in the Milky Way contain black holes. The reason for this was explained by Sigurdsson and Hernquist (1993) and Kulkarni et al. (1993). If a significant population of stellar-mass black holes is retained in a cluster, the black holes will sink by mass segregation forming their own sub-cluster – a *black cluster* – at the cluster center via the so-called Spitzer instability (Spitzer 1987, and also (● 17.41) of this chapter). These black holes will then interact with each other, rapidly forming binaries. The black-hole population will be depleted as black holes will be ejected via binary-single scattering.

The black cluster formed at the heart of a stellar cluster will also have an impact on the cluster's dynamical evolution (Mackey et al. 2007, 2008). If black holes are ejected from a cluster, the cluster remaining will expand slightly due to the reduction in binding gravitational mass. More effectively, black holes ejected from the black cluster but only into the halo of the cluster will then sink back into the core again transferring energy to the cluster stars in the process. Thus the cluster will be heated and expand, both at the core and half-mass radii. The effectiveness and time dependence of this heating process will depend chiefly on two factors: the retention fraction of the black holes produced in core-collapse supernovae, and the degree of initial mass segregation of the massive stars which produce the stellar-mass black holes.

Nonsegregated systems with a low black-hole retention fraction will evolve toward core collapse as described in (● Sect. 3). Those retaining a sizeable population of black holes will experience significant core expansion due to heating from the black cluster but only after about one billion years, after the black core has been established, black-hole binaries have formed, and ejection of black holes in to the halo and out of the cluster has begun.

Segregated systems with a low black-hole retention fraction will also experience core expansion, but much earlier. This is because mass loss due to stellar evolution (i.e., the massive stars exploding as supernovae) will have a greater effect as now all the massive stars are located deep in the potential well of the cluster owing to their segregation (see Vesperini et al. 2009). Segregated systems retaining their black holes will in addition experience later heating of the core in a manner similar to the nonsegregated case.

Thus clusters today will have a broad range of core radii, as is observed, which is illustrated in (● Fig. 17-11 (taken from Mackey et al. 2008)). In this figure, the core radii and cluster ages for the rich stellar clusters observed in the LMC and SMC are shown. One can see how the observed clusters can be explained as coming from clusters with or without mass segregation and black-hole retention.



■ Fig. 17-11

Core radius [pc] plotted as a function of $\log(\text{time})$ for several model clusters with various assumptions about the *black hole* population retained and the degree of initial mass segregation. In runs 1, 2, and 6 there was no initial mass segregation. In runs 2, 4, 4a, and 4b, the *black holes* produced in core-collapse supernovae were retained in the cluster. Runs 4, 4a and 4b differ in the degree of initial mass segregation. The data points plotted along with their error bars are the observed LMC and SMC clusters (Fig. 28 of Mackey et al. (2008), reproduced with permission)

4.3.2 Observational Evidence for Intermediate-Mass Black Holes

There is extremely good observational evidence for the existence of stellar mass black holes from X-ray binaries, and for supermassive black holes from dynamical observations of galactic nuclei. However, there is a relative lack of evidence for the existence of black holes of intermediate mass (between 100 and $10^5 M_{\odot}$). Extrapolating from the observed relationship between black hole mass and velocity dispersion seen in the bulges of galactic nuclei would suggest that globular clusters could host such intermediate-mass black holes (IMBHs).

One might imagine that the presence of an IMBH within the core of a cluster would be revealed by X-ray emission due to the accretion of material into it. A problem here, however, is that an ensemble of lower-luminosity X-ray objects, such as LMXBs and CVs, may together appear as a single brighter source, if the cluster is not resolved sufficiently.

An alternative approach is to compare dynamical modeling of clusters with observed radial kinematics. It has been understood for a long time that the presence of an IMBH will affect the stellar orbits at a cluster center (Bahcall and Wolf 1976), producing a cuspy density profile. IMBHs have been suggested to be present in three clusters: M15 and ω Centauri in our own galaxy (Gerssen et al. 2002; Noyola et al. 2008) and G1 in M31 (Gebhardt et al. 2002, 2005).

However other work questions the need for IMBHs in M15 and G1 (Baumgardt et al. 2003; McNamara et al. 2003), with N-body simulations of clusters providing good fits to the observations without the need for an IMBH. One should also recall that G1 and ω Centauri are rather unusual globular clusters. Both are rather massive and could be the nuclei of former dwarf galaxies.

Baumgardt et al. (2004a, b) performed N-body simulations of stellar clusters which also contained a single $1,000 M_{\odot}$ black hole. They found that rather than having cuspy profiles reminiscent of a post-core-collapse cluster, the surface density of bright stars is rather flat as the cores are heated via two-body scattering and mass loss due to stellar evolution. Stars close to the black hole transfer energy to more-distant stars, leaving themselves more deeply bound until they are tidally disrupted by the central black hole. Thus clusters containing IMBHs could more easily be fit with King models (Baumgardt et al. 2005). It has also been demonstrated by using N-body simulations of clusters that the presence of an IMBH may prevent, or at least slow down, the process of mass segregation within a cluster (Gill et al. 2008). Thus one might be able to infer the presence, or absence, of an IMBH within a cluster by measuring the degree of mass segregation for dynamically evolved clusters (i.e., clusters with a half-mass relaxation time less than 1 Gyr).

4.3.3 Producing Intermediate-Mass Black Holes in Clusters

In this subsection, two potential ways of producing IMBHs within stellar clusters will be considered. By the first route, stellar collisions occur frequently enough that the time between collisions becomes shorter than the stellar lifetimes. This can be the case if the core of the cluster reaches very high number densities of stars sufficiently early enough in the evolution of the cluster (i.e., before the massive stars have a chance to evolve and explode as supernovae). In such a circumstance, it is at least theoretically possible to build up very massive stars, which could potentially produce IMBHs when they undergo core-collapse, providing certain conditions are met: (1) it is necessary that collisions result in stellar mergers where a relatively small fraction of the mass is ejected, (2) collision products need to be well mixed in order to provide fresh fuel for the fusion reactions in the stellar core, and (3) the merger product must be able to form a stable star.

Work by Portegies Zwart and McMillan (2002) and Portegies Zwart et al. (2004) showed that the timescale for collisions could become very short provided the clusters formed with very short half-mass relaxation timescales so the most massive stars will have sunk into the core before they explode as supernovae. Once there, they encounter (and merge with) other massive stars to potentially form a very massive star which could then be a progenitor for an IMBH. Both the work cited above and Monte Carlo calculations (Gürkan et al. 2004) show that about 0.1% of the cluster mass could end up in this very massive object.

However, even if the initial conditions of a cluster are such that the massive stars can sink rapidly enough, forming a core containing massive stars of sufficient density for the collision timescale to sink below the evolutionary timescale, there is another barrier to mass growth through collisions: mass loss through stellar winds. Once a merger product becomes reasonably massive (about $50 M_{\odot}$), mass loss through winds becomes significant even at earlier evolutionary stages. Detailed calculations of the evolution of post-merger objects show that mass loss from winds prevents the stellar mass from growing (Glebbeek et al. 2009). In other

words, runaway mergers of main-sequence stars cannot produce a very massive star capable of producing an IMBH via a core-collapse supernova.

An alternative route is to form lower-mass black holes first, through the core-collapse supernovae of several stars, and then to have these black holes sink to the cluster core where they can pick up mass by either merging with other black holes or accreting material from stars. In particular, the relatively massive black holes are likely to exchange into any binaries present in the cluster core. This process will repeat, producing many binaries containing two black holes, which could spiral-in and merge via the action of gravitational radiation provided the binaries are sufficiently tight. If mass is built up predominantly via black-hole mergers, the merging system has to avoid cluster ejection via the gravitational radiation recoil kick binaries experience when they merge. This kick can exceed several 100 km/s (Baker et al. 2008; Redmount and Rees 1989). The recoil of the system will be reduced for the merger of two black holes of very different masses. In other words, there is a certain minimum mass required of the more massive black hole in the binary, given that the lower-mass secondaries are likely to have masses around $10 M_{\odot}$ having been produced directly in core-collapse supernovae. Miller and Hamilton (2002) suggest that IMBHs may be relatively common in the centers of dense globular clusters, providing that they contained at least one black hole of mass $\geq 50 M_{\odot}$. Later calculations suggest that an initial minimum black-hole mass of about $400 M_{\odot}$ is required to guarantee cluster retention when this black hole merges with a $10 M_{\odot}$ black hole (Baker et al. 2008). Both Holley-Bockelmann et al. (2008) and Moody and Sigurdsson (2009) suggest that in the vast majority of cases, black-hole binaries are most likely to get ejected from their host clusters, and only a few clusters could contain IMBHs made in this way today.

4.4 Multiple Stellar Populations

Thus far, globular clusters have been portrayed as relatively simple objects. The discussions of dynamical evolution in the earlier sections assumed that the cluster contained a single generation of stars, that any remaining gas was ejected early on in the cluster's history, and that there was no subsequent star formation. Detailed observations of some clusters, such as NGC 6397, support this picture. However, recent observations have revealed a more complex reality for several clusters. The observational evidence for multiple stellar populations comes in three distinct forms: (1) multiple main sequences seen in color-magnitude diagrams derived from photometric observations of clusters; (2) multiple turn-off and sub-giant branches, again seen through photometric observations; and (3) anti-correlations in the abundances of light elements such as sodium and oxygen which suggest that the stars have formed from gas which has been involved in nuclear reactions in an earlier generation of stars.

The first cluster seen to contain multiple main sequences was ω Centauri, where two main sequences were first observed (Bedin et al. 2004). Not only does this cluster contain two main sequences, spectroscopic analysis has shown the bluer main sequence to be extremely metal rich and enriched in helium ($Y \sim 0.38$) relative to the second stellar population (Piotto et al. 2005). This is a huge amount of enrichment. Of course, ω Centauri is already an unusual object, being one of the most massive clusters. It could be the nucleus of a former dwarf galaxy.

However, ω Centauri is not unique in showing multiple main sequences. The color-magnitude diagram of the globular cluster NGC 2808 is split into three main sequences

(Piotto et al. 2007). Stars in this cluster also seem to have a spread in helium abundances (D'Antona et al. 2005). Spectroscopy of red giants reveals an anticorrelation in sodium and oxygen abundances (Carretta et al. 2006). Such an anticorrelation occurs in material taking part in hot H-burning. The picture then emerges that these stars have been formed from material processed in a previous generation of stars.

Observations reveal splits in the turn off and sub-giant branches in several clusters in our galaxy and also in the LMC (Piotto 2009). It would seem that multiple populations are phenomena common to a large fraction of clusters. It is unclear whether the same processes are responsible for the split main sequences, sub-giant branches and for the abundance anticorrelations seen on the red-giant branch. They do all indicate that chemical enrichment and multiple periods of star formation occur in at least some globular clusters.

Hot hydrogen burning is thought to produce the helium-enriched material, and also to give at the same time the observed oxygen-sodium abundance anticorrelation. This requires that the temperature where the nuclear burning took place is sufficiently high, and that the enriched material can be retained by the cluster to form the second generation of stars. Possible sources of the enriched gas include intermediate-mass asymptotic giant branch (AGB) stars (Ventura et al. 2001) and massive ($M > 10 M_{\odot}$) rapidly rotating stars which are likely to have very strong winds (Decressin et al. 2007). Alternatively, massive stars within tight binaries could also lose large amounts of chemically enriched mass (de Mink et al. 2009). The key idea here is that mass can be lost in interacting binaries in non-conservative mass transfer. Considering a Kroupa initial distribution function (IMF), de Mink et al. argue that mass loss from massive interacting binaries will be larger than from AGB stars and very massive rapidly rotating stars combined.

The formation and dynamical evolution of multiple generations of stars has been studied theoretically using a combination of hydrodynamic and N-body simulations (D'Ercole et al. 2008). Using a spherically symmetric hydrodynamic simulation, they find that gas ejected from a first generation of AGB stars collects in a cooling flow in the cluster core where it forms a second generation of stars. They also consider the case where unpolluted gas is mixed in with the AGB wind material. Using N-body simulations, they show how a large fraction of the first generation of stars are lost when the cluster expands due to mass loss from supernovae. A much larger fraction of the second generation of stars are retained as they are mostly located in the cluster core. Thus it is possible for the second generation to outnumber the first in the cluster today. In such a scenario, the clusters seen today have lost the majority of their stars: it could be then that some clusters have initial masses as large as $10^8 M_{\odot}$.

5 Cluster Survival

In this final section the role played by our galaxy in the evolution of globular clusters is considered. In particular how the rate of mass loss from clusters is enhanced by the galactic tidal field is considered. The stellar contents of a globular cluster will be stirred up as it passes through the galactic disk or close to the bulge in a process known as disk and bulge shocking. Finally dynamical friction will cause any globular clusters formed relatively close to the galactic center to spiral in and be lost. By considering all three processes together with the internal evolution of clusters driven by two-body relaxation, one is able to investigate how the entire population of galactic globular clusters evolves over time. Low-mass clusters and clusters closer to the galactic center will be vulnerable to destruction.

5.1 Stellar Evaporation and Tidal Truncation

As has been seen earlier, energy flows between stars in a cluster via two-body scattering. One can imagine a situation where a single strong encounter between two stars leads to the ejection of one star from the cluster as its speed exceeds the local escape speed of the cluster. Following the nomenclature of Binney and Tremaine (2008), I will refer to such a process as stellar ejection. Alternatively, one could imagine a series of several, weaker, encounters where a star gradually increases its energy until it has a small positive energy and escapes. One can label this process as evaporation (Binney and Tremaine 2008).

The ejection rate for an isolated cluster with a Plummer density distribution was calculated by Hénon (1961, 1969). For a cluster containing stars of a single mass, the ejection rate is given by

$$\frac{dN}{dt} = -1.05 \times 10^{-3} \frac{N}{t_{\text{rh}} \ln(\Lambda N)} \quad (17.51)$$

where t_{rh} is the half-mass relaxation time. Thus an ejection timescale is given by

$$t_{\text{eject}} = - \left(\frac{1}{N} \frac{dN}{dt} \right)^{-1} = 1 \times 10^3 \ln(\Lambda N) t_{\text{rh}} \quad (17.52)$$

Typically $\ln(\Lambda N) \sim 10$ and t_{eject} will be much longer than the evaporation timescale. Fokker–Planck calculations show that the evaporation timescale for an isolated cluster composed of single-mass stars is given by (Spitzer 1987)

$$t_{\text{evap}} = - \left(\frac{1}{N} \frac{dN}{dt} \right)^{-1} = f t_{\text{rh}} \quad (17.53)$$

where $f \simeq 300$. In clusters having a range of stellar masses, low-mass stars will tend to evaporate more quickly as the effects of mass segregation will redistribute them further out in the cluster (Giersz and Heggie 1996; Chernoff and Weinberg 1990). As has been seen earlier, core collapse occurs after 10–20 t_{rh} , thus a very small fraction of stars will evaporate before core collapse. An isolated cluster expands after core collapse leading to increases in the half-mass relaxation timescale and the evaporation timescale. Isolated clusters would therefore take an extremely long time (much longer than the age of the universe) to lose the majority of their stars (Baumgardt et al. 2002).

The presence of the galactic tidal field greatly accelerates stellar evaporation rates from globular clusters (e.g., Giersz and Heggie 1997; Baumgardt and Makino 2003). Fokker–Planck calculations have been used to calculate the coefficient f in (17.53) for globular clusters in the tidal field of our galaxy. It is found that $f \simeq 20 - 60$ (Spitzer 1987). For a cluster in the tidal field of the galaxy, the evaporation rate is increased once the cluster expands after core collapse and bounce, as it overflows its tidal radius by a larger amount.

5.2 Disk and Bulge Shocking

Time-varying tidal forces lead to gravitational shocking as a globular cluster passes through the galactic disk (Ostriker et al. 1972) or passes by the galactic bulge (Spitzer 1987). Tidal shocks heat a stellar population in the outer regions of a globular cluster, enhancing mass loss

(e.g., Dehnen et al. 2004; Grillmair et al. 1995; Leon et al. 2000). Cluster evolution can be accelerated (Gnedin et al. 1999). The key idea here is that the timescale for a cluster to pass through the disk will be much shorter than the orbital timescale for stars in the outer parts of the cluster (hence the use of the word “shock”).

One can therefore employ the impulse approximation to get some understanding of the effect of tidal shocks and to see how the timescale for mass loss due to tidal shocking scales with cluster properties. Here the impulse approximation applied to disk shocks following the approach described in Heggie and Hut (2003) is considered. Consider a galactic disk modeled as an (infinite) sheet of matter of surface density Σ . The acceleration due to the disk is then $2\pi G\Sigma$. The acceleration is directed downward for objects above the plane of the disk, and upward for those located below the disk. A star some distance r from the center of the cluster will experience a differential acceleration with respect to the cluster center of $4\pi G\Sigma r$ for a time of order r/V where V is the orbital velocity of the stellar cluster, as for this time the star and cluster center will be located on different sides (above and below) the disk. The speed of the star in the frame of the cluster is then changed by an amount $\delta v \sim 4\pi G\Sigma r/V$. The kinetic energy of the entire cluster is increased by an amount given approximately by $M_c(\delta v)^2/2 \sim 8\pi^2 G^2 M_c \Sigma^2 r^2/V^2$, where M_c is the mass of the cluster.

The time between tidal shocks is approximately $\pi R/V$ where R is the radius of the cluster orbit in the galaxy. The time for shocks to destroy the cluster, t_{sh} , will be roughly given by the timescale for disk shocks to double the kinetic energy of the cluster. Thus

$$t_{\text{sh}} = \left(\frac{v}{\delta v} \right)^2 \frac{\pi R}{V} = \frac{v^2 R V}{16\pi G^2 \Sigma^2 r^2} \quad (17.54)$$

For a virialized cluster, $v^2 \simeq 0.45 GM_c/r_h$ where r_h is the half-mass radius (Binney and Tremaine 2008). Thus

$$t_{\text{sh}} \simeq 10^{-2} \frac{M_c R V}{G \Sigma^2 r_h^3} \quad (17.55)$$

More detailed calculations reveal that correction factors have to be employed to avoid overestimating the effect (Gnedin and Ostriker 1997), although the scaling with cluster properties derived above remain.

5.3 Inspiral Due to Dynamical Friction

As a globular cluster passes through the halo, material will be swept up behind it via gravitational deflections. This excess of material behind it will exert a net gravitational force – a drag – on the cluster. This force will be directed in the direction opposite to the velocity of the cluster and will cause it to lose angular momentum and spiral toward the center of the galaxy.

The analysis of Binney and Tremaine (1987, 2008) is outlined below. The acceleration on a globular cluster of mass M_c passing through a sea of stars, each of mass m , is given by

$$\frac{d\mathbf{v}_{M_c}}{dt} = -16\pi^2 \ln\Lambda G^2 m M_c \frac{\int_0^{v_M} f(v_m) v_m^2 dv_m}{v_{M_c}^3} \mathbf{v}_{M_c} \quad (17.56)$$

where, as earlier, $\ln\Lambda$ is the Coulomb logarithm (see [Sect. 3.2](#)), and $f(v_m)$ is the distribution function for the stars. The above formula is known as the Chandrasekhar dynamical friction

formula. One can see from the above equation that only stars moving more slowly than the globular cluster exert a force. If v_{M_c} is small, one may use $f(0)$ rather than perform the integral in the above equation and the drag force $\dot{\mathbf{v}}_{M_c} = -K\mathbf{v}_{M_c}$, where K is some constant. For much larger values of v_{M_c} , the integral approaches a constant and the frictional force is proportional to $v_{M_c}^2$. If $f(v_m)$ follows a Maxwellian distribution function with a dispersion σ then

$$\frac{d\mathbf{v}_{M_c}}{dt} = -\frac{4\pi\ln\Lambda G^2\rho M_c}{v_{M_c}^3} \left[\operatorname{erf}(X) - \frac{2X}{\sqrt{\pi}} e^{-X^2} \right] \mathbf{v}_{M_c} \quad (17.57)$$

where erf , the so-called error function, is given by $\operatorname{erf}(X) = (2/\sqrt{\pi}) \int_0^X e^{-x^2} dx$.

Equation (17.57) can now be applied to the problem of globular cluster inspiral within our galaxy. As the rotation curve for our galaxy is rather flat, the density may be approximated by that of the singular isothermal sphere (see Sect. 2). After some algebra one finds that the frictional force on a globular cluster becomes $F = -0.428\ln\Lambda GM_c^2/R^2$, where R is the galactocentric distance of the globular cluster. This frictional force will cause the cluster to lose angular momentum at a rate $\dot{L} = FR/M_c$. If one assumes the inspiral rate is relatively slow, and that the cluster remains on an almost circular orbits, and thus $L = RV_c$, where V_c is the orbital speed of the globular cluster, then one has

$$R \frac{dR}{dt} = -0.428 \frac{GM_c}{V_c} \ln\Lambda \quad (17.58)$$

Considering a cluster beginning at a radius R_i , one may solve the above differential equation to compute the time required for a globular cluster on a circular orbit to spiral in to the galactic center from some initial radius R_i as

$$t_{\text{fric}} = \frac{1.17 R_i^2 V_c}{\ln\Lambda GM_c} = \frac{2.6 \times 10^{11}}{\ln\Lambda} \left(\frac{R_i}{2 \text{ kpc}} \right)^2 \left(\frac{V_c}{250 \text{ km/s}} \right) \left(\frac{10^6 M_\odot}{M_c} \right) \text{ year} \quad (17.59)$$

Thus considering typical globular clusters, only those initially within 2 kpc or so of the galactic center will spiral in to the center on interesting timescales.

5.4 The Combined Effect on the Cluster Population

All the elements can now be put together in order to model the survivability of globular clusters within our galaxy. Three effects will independently act to destroy globular clusters: (1) mass loss due to stellar evaporation, accelerated by the presence of the galactic tidal field, (2) tidal shocks acting on a cluster whenever it passes through the galactic disk or near to the bulge, and (3) dynamical friction, which will act to drag a cluster inward in a spiral motion toward the galactic center.

In Sect. 5.1, the timescale for mass loss through evaporation was shown to be $t_{\text{evap}} \propto t_{\text{rh}} \propto M_c^{1/2} r_h^{3/2}$, where t_{rh} is the half-mass radius and M_c is the cluster mass. If the evaporation timescale is set to the age of the universe, then one sees that this will be true along a locus of points where $M_c^{1/2} r_h^{3/2} = \text{constant}$, or in other words, where $M_c \propto r_h^{-3}$. Evaporation will only be important for more massive clusters having smaller half-mass radii. The timescale for this process will also be independent of the stellar cluster position within the galaxy.

In [Sect. 5.2](#), the timescale for cluster dissolution via tidal shocking was shown to be given by $t_{\text{sh}} \propto M_c R V / G \Sigma^2 r_h^3$, where R is the galactic radius of the stellar-cluster orbit, V is the cluster orbital speed, and Σ is the surface density of the galactic disk which the cluster plunges through periodically while on its orbit. The effects of tidal shocking are therefore dependent on the position of the cluster in the galaxy. For a given galactocentric radius, and again setting the timescale to the age of the universe, then $M_c \propto r_h^3$. More massive clusters will be vulnerable to destruction by disk and bulge shocks if they have larger half-mass radii.

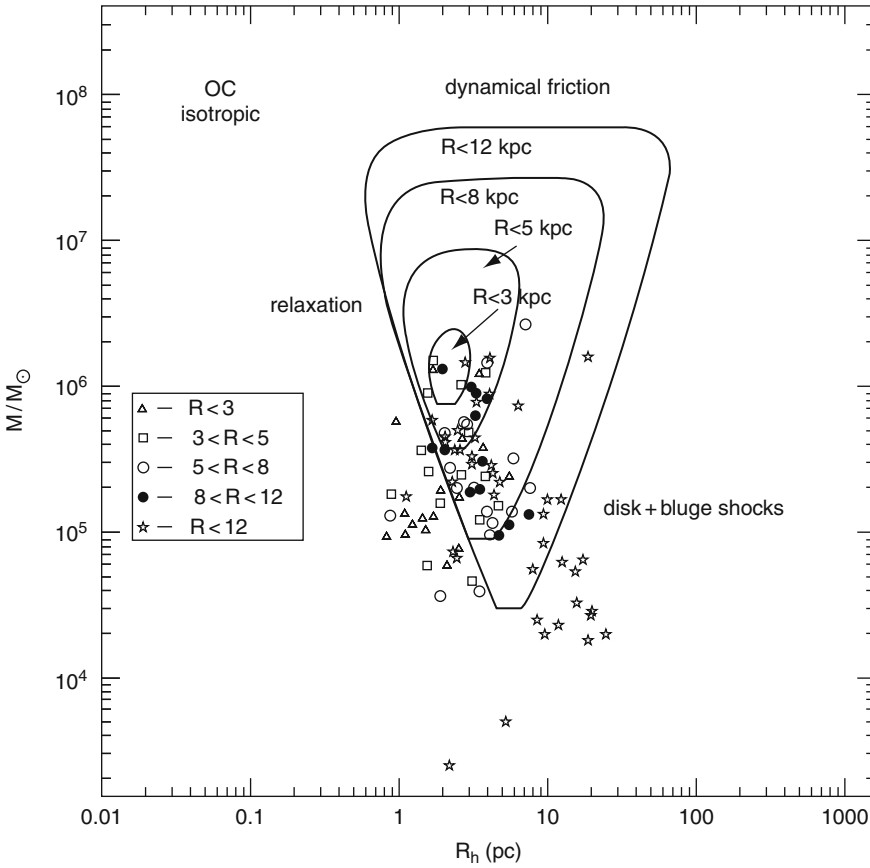
In [Sect. 5.3](#), the timescale for a cluster to sink into the galactic center via dynamical friction was shown to be given by $t_{\text{fric}} \propto M_c^{-1} R_i^2 V_c$, where R_i is the initial galactocentric radius of the cluster, and V_c is the orbital speed. The effects of dynamical friction do not depend on the cluster half-mass radius. Fixing the dynamical friction timescale, more massive clusters will be able to sink in to the center from greater initial distances, where the cluster mass and maximum initial galactocentric radius are related by $M_c \propto R_i^2$.

One can combine the effects of all three processes and consider the locus of points in the $M_c - r_h$ plane where the timescale for the total destruction rate is the Hubble time (Aguilar et al. 1988; Gnedin and Ostriker 1997):

$$\frac{1}{t_{\text{Hubble}}} = \frac{1}{t_{\text{evap}}} + \frac{1}{t_{\text{sh}}} + \frac{1}{t_{\text{fric}}} \quad (17.60)$$

In [Fig. 17-12](#), [Fig. 21](#) of Gnedin and Ostriker (1997) is reproduced, where the effects of all three destruction processes are shown in the $M_c - r_h$ plane for various galactocentric radii, together with the observed population of globular clusters. The region where clusters can survive forms a triangle in the $M_c - r_h$ plane. The locations of observed clusters suggest that both evaporation and disk and bulge shocks play a role in shaping the globular cluster population. The latter would suggest that properties of clusters would correlate with their position in the galaxy. This is seen, for example, clusters closer to the galactic center are observed to have higher concentrations and have smaller and denser cores (Djorgovski and Meylan 1994).

Gnedin and Ostriker (1997) performed Fokker–Planck calculations of cluster evolution including the effects described above, with some assumptions being made about the globular cluster orbits (as not all clusters have measured radial velocities and proper motions). Their results suggest that 50–90% of the current globular cluster population will be destroyed in the next 10^{10} years. This destruction rate is about a factor of 10 higher than predicted earlier simply using timescale arguments (Aguilar et al. 1988). Dinescu et al. (1999) followed the approach of Gnedin and Ostriker, but for a subset of 38 clusters for which both the radial velocities and proper motions were known. In other words, in their sample, the actual cluster orbits are known. They found that internal relaxation and evaporation were more important than disk shocking for these clusters. Vesperini and Heggie (1997) performed N-body simulations of globular clusters including the effects of two-body relaxation, stellar evolution, disk shocking, and galactic tidal fields, producing expressions for the time evolution of cluster mass as a function of initial cluster mass and galactocentric radius (orbits were assumed to be circular). These expressions were then used by Vesperini (1998) to model the evolution of a population of globular clusters. Similar results to Gnedin and Ostriker were found, namely that about 50% of the initial clusters would be destroyed by today. Other calculations point out that low-mass clusters are particularly vulnerable to destruction (e.g., Baumgardt 1998) with the mass function for globular clusters developing a peak at later times around $2 \times 10^5 M_\odot$ for a wide variety of initial cluster populations (Fall and Zhang 2001).



■ Fig. 17-12

The contributions made to the destruction of globular clusters by three mechanisms: stellar evaporation (labeled here as relaxation), tidal shocks, and dynamical friction (Fig. 21 of Gnedin and Ostriker (1997), reproduced with permission). OC refers to the model for the galaxy adopted, and isotropic to the assumptions made about the clusters' orbits within the galaxy. The lines show the locus of points where the combined destruction rates for all three mechanisms give a destruction timescale of a Hubble time. Lines are plotted for various galactocentric radii

One is led to conclude from all these studies that the globular cluster population is shaped significantly by internal dynamical evolution and by our galaxy on the timescale of 10^{10} years: the globular cluster population was almost certainly much larger in the past, and will be much reduced in the future.

Acknowledgments

I thank the following people for their input and comments on earlier versions of this chapter: Ross Church, Sofia Feltzing, Douglas Heggie, Berry Holl, Lennart Lindegren, Steve McMillan, and Giampaolo Piotto.

Cross-References

- [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- [Dynamics of Disks and Warps](#)
- [Mass Distribution and Rotation Curve in the Galaxy](#)

References

- Aarseth, S. J. 1999, *PASP*, 111, 1333
- Aarseth, S. J. 2003, in *Gravitational N-Body Simulations*, ed. S. J. Aarseth
- Aarseth, S. J., Tout, C. A., & Mardling, R. A. (eds) 2008, *The Cambridge N-Body Lectures, Lecture Notes in Physics*, Vol. 760
- Aguilar, L., Hut, P., & Ostriker, J. P. 1988, *ApJ*, 335, 720
- Amaro-Seoane, P., Gair, J. R., Freitag, M., et al. 2007, *Classical Quant Grav*, 24, 113
- Antonov, V. A. 1962, in *Solution of the Problem of Stability of Stellar System Emden's Density Law and the Spherical Distribution of Velocities*, ed. V. A. Antonov
- Ashman, K. M., & Zepf, S. E. 1998, in *Globular Cluster Systems*, ed. K. M. Ashman, & S. E. Zepf
- Bahcall, J. N., & Wolf, R. A. 1976, *ApJ*, 209, 214
- Bailyn, C. D. 1995, *ARA&A*, 33, 133
- Baker, J. G., Boggs, W. D., Centrella, J. et al. 2008, *Astrophys J Lett*, 682, L29
- Baumgardt, H. 1998, *A&A*, 330, 480
- Baumgardt, H., & Kroupa, P. 2007, *MNRAS*, 380, 1589
- Baumgardt, H., & Makino, J. 2003, *MNRAS*, 340, 227
- Baumgardt, H., Hut, P., & Heggie, D. C. 2002, *MNRAS*, 336, 1069
- Baumgardt, H., Makino, J., Hut, P., McMillan, S., & Portegies Zwart, S. 2003, *Astrophys J Lett*, 589, L25
- Baumgardt, H., Makino, J., & Ebisuzaki, T. 2004a, *ApJ*, 613, 1133
- Baumgardt, H., Makino, J., & Ebisuzaki, T. 2004b, *ApJ*, 613, 1143
- Baumgardt, H., Makino, J., & Hut, P. 2005, *ApJ*, 620, 238
- Bedin, L. R., Piotto, G., Anderson, J., et al. 2004, *Astrophys J Lett*, 605, L125
- Benz, W., & Hills, J. G. 1987, *ApJ*, 323, 614
- Benz, W., & Hills, J. G. 1992, *ApJ*, 389, 546
- Binney, J., & Tremaine, S. 1987, in *Galactic Dynamics*, ed. J. Binney, & S. Tremaine
- Binney, J., & Tremaine, S. 2008, *Galactic Dynamics: Second Edition*, ed. J. Binney, & S. Tremaine (Princeton: Princeton University Press)
- Bonnell, I. A., & Davies, M. B. 1998, *MNRAS*, 295, 691
- Carretta, E., Bragaglia, A., Gratton, R. G., et al. 2006, *A&A*, 450, 523
- Chandrasekhar, S. 1967, *An introduction to the study of stellar structure*, ed. S. Chandrasekhar
- Chernoff, D. F., & Weinberg, M. D. 1990, *ApJ*, 351, 121
- Cohn, H. 1979, *ApJ*, 242, 1036
- Cohn, H. 1980, *ApJ*, 242, 765
- D'Antona, F., Bellazzini, M., Caloi, V., et al. 2005, *ApJ*, 631, 868
- Davies, M. B. 1995, *MNRAS*, 276, 887
- Davies, M. B. 1997, *MNRAS*, 288, 117
- Davies, M. B., & Benz, W. 1995, *MNRAS*, 276, 876
- Davies, M. B., Benz, W., & Hills, J. G. 1991, *ApJ*, 381, 449
- Davies, M. B., Piotto, G., & de Angeli, F. 2004, *MNRAS*, 349, 129
- de Mink, S. E., Pols, O. R., Langer, N., & Izzard, R. G. 2009, *A&A*, 507, L1
- Decressin, T., Meynet, G., Charbonnel, C., Prantzos, N., & Ekström, S. 2007, *A&A*, 464, 1029
- Dehnen, W., Odenkirchen, M., Grebel, E. K., & Rix, H. 2004, *AJ*, 127, 2753
- D'Ercole, A., Vesperini, E., D'Antona, F., McMillan, S. L. W., & Recchi, S. 2008, *MNRAS*, 391, 825
- Dinescu, D. I., Girard, T. M., & van Altena, W. F. 1999, *AJ*, 117, 1792
- Djorgovski, S., & Meylan, G. 1994, *AJ*, 108, 1292
- Drukier, G. A. 1995, *ApJS*, 100, 347
- Elson, R. A. W., Fall, S. M., & Freeman, K. C. 1987, *ApJ*, 323, 54
- Fabian, A. C., Pringle, J. E., & Rees, M. J. 1975, *MNRAS*, 172, 15P
- Fall, S. M., & Zhang, Q. 2001, *ApJ*, 561, 751
- Fregeau, J. M., & Rasio, F. A. 2007, *ApJ*, 658, 1047
- Fregeau, J. M., Gürkan, M. A., Joshi, K. J., & Rasio, F. A. 2003, *ApJ*, 593, 772
- Fukushige, T., & Heggie, D. C. 1995, *MNRAS*, 276, 206
- Fukushige, T., & Heggie, D. C. 2000, *MNRAS*, 318, 753

- Gao, B., Goodman, J., Cohn, H., & Murphy, B. 1991, *ApJ*, 370, 567
- Gebhardt, K., Rich, R. M., & Ho, L. C. 2002, *Astrophys J Lett*, 578, L41
- Gebhardt, K., Rich, R. M., & Ho, L. C. 2005, *ApJ*, 634, 1093
- Gerssen, J., van der Marel, R. P., Gebhardt, K., et al. 2002, *AJ*, 124, 3270
- Giersz, M., & Heggie, D. C. 1996, *MNRAS*, 279, 1037
- Giersz, M., & Heggie, D. C. 1997, *MNRAS*, 286, 709
- Giersz, M., & Heggie, D. C. 2009, *MNRAS*, 395, 1173
- Gill, M., Trenti, M., Miller, M. C., et al. 2008, *ApJ*, 686, 303
- Glebbeeck, E., Gaburov, E., de Mink, S. E., Pols, O. R., & Portegies Zwart, S. F. 2009, *A&A*, 497, 255
- Gnedin, N. Y., & Ostriker, J. P. 1997, *ApJ*, 486, 581
- Gnedin, O. Y., Lee, H. M., & Ostriker, J. P. 1999, *ApJ*, 522, 935
- Goodman, J. 1984, *ApJ*, 280, 298
- Gratton, R., Sneden, C., & Carretta, E. 2004, *ARA&A*, 42, 385
- Grillmair, C. J., Freeman, K. C., Irwin, M., & Quinn, P. J. 1995, *AJ*, 109, 2553
- Gunn, J. E., & Griffin, R. F. 1979, *AJ*, 84, 752
- Gürkan, M. A., Freitag, M., & Rasio, F. A. 2004, *ApJ*, 604, 632
- Heggie, D. C. 1975, *MNRAS*, 173, 729
- Heggie, D. C., & Giersz, M. 2008, *MNRAS*, 389, 1858
- Heggie, D., & Hut, P. 2003, *The Gravitational Million-Body Problem: A Multidisciplinary Approach to Star Cluster Dynamics*, ed. D. Heggie, & P. Hut
- Heggie, D. C., Trenti, M., & Hut, P. 2006, *MNRAS*, 368, 677
- Hénon, M. 1961, *Annales d'Astrophysique*, 24, 369
- Hénon, M. 1964, *Annales d'Astrophysique*, 27, 83
- Henon, M. 1969, *A&A*, 2, 151
- Hénon, M. 1971, *Ap&ss*, 14, 151
- Hills, J. G. 1975, *AJ*, 80, 809
- Hills, J. G., & Day, C. A. 1976, *Astrophys Lett*, 17, 87
- Holley-Bockelmann, K., Gültekin, K., Shoemaker, D., & Yunes, N. 2008, *ApJ*, 686, 829
- Hunter, C. 2001, *MNRAS*, 328, 839
- Hurley, J. R. 2007, *MNRAS*, 379, 93
- Hurley, J. R., Tout, C. A., Aarseth, S. J., & Pols, O. R. 2001, *MNRAS*, 323, 630
- Hurley, J. R., Pols, O. R., Aarseth, S. J., & Tout, C. A. 2005, *MNRAS*, 363, 293
- Hurley, J. R., Aarseth, S. J., & Shara, M. M. 2007, *ApJ*, 665, 707
- Hut, P., & Bahcall, J. N. 1983, *ApJ*, 268, 319
- Hut, P., McMillan, S., Goodman, J., et al. 1992a, *PASP*, 104, 981
- Hut, P., McMillan, S., & Romani, R. W. 1992b, *ApJ*, 389, 527
- Ivanova, N., Belczynski, K., Fregeau, J. M., & Rasio, F. A. 2005, *MNRAS*, 358, 572
- Ivanova, N., Heinke, C. O., Rasio, F. A., et al. 2006, *MNRAS*, 372, 1043
- Ivanova, N., Heinke, C. O., Rasio, F. A., Belczynski, K., & Fregeau, J. M. 2008, *MNRAS*, 386, 553
- King, I. 1962, *AJ*, 67, 471
- King, I. R. 1966, *AJ*, 71, 64
- King, I. R., Sosin, C., & Cool, A. M. 1995, *Astrophys J Lett*, 452, L33+
- King, I. R., Cool, A. M., & Piotto, G. 1996, in *Formation of the Galactic Halo...Inside and Out*, *Astronomical Society of the Pacific Conference Series*, Vol. 92, ed. H. L. Morrison, & A. Sarajedini 277+–
- Kroupa, P., Aarseth, S., & Hurley, J. 2001, *MNRAS*, 321, 699
- Kulkarni, S. R., Hut, P., & McMillan, S. 1993, *Nature*, 364, 421
- Larson, R. B. 1970a, *MNRAS*, 147, 323
- Larson, R. B. 1970b, *MNRAS*, 150, 93
- Lee, H. M., & Ostriker, J. P. 1986, *ApJ*, 310, 176
- Leon, S., Meylan, G., & Combes, F. 2000, *A&A*, 359, 907
- Lynden-Bell, D. 1962, *MNRAS*, 124, 279
- Lynden-Bell, D. 1967, *MNRAS*, 136, 101
- Lynden-Bell, D., & Eggleton, P. P. 1980, *MNRAS*, 191, 483
- Lynden-Bell, D., & Wood, R. 1968, *MNRAS*, 138, 495
- Mackey, A. D., Wilkinson, M. I., Davies, M. B., & Gilmore, G. F. 2007, *MNRAS*, 379, L40
- Mackey, A. D., Wilkinson, M. I., Davies, M. B., & Gilmore, G. F. 2008, *MNRAS*, 386, 65
- Makino, J. 1996, *ApJ*, 471, 796
- McMillan, S. L. W., McDermott, P. N., & Taam, R. E. 1987, *ApJ*, 318, 261
- McNamara, B. J., Harrison, T. E., & Anderson, J. 2003, *ApJ*, 595, 187
- Meylan, G. 1987, *A&A*, 184, 144
- Meylan, G. 1988, *A&A*, 191, 215
- Meylan, G., & Heggie, D. C. 1997, *Astron Astrophys Rev*, 8, 1
- Meylan, G., & Mayor, M. 1986, *A&A*, 166, 122
- Michie, R. W. 1963, *MNRAS*, 125, 127
- Miller, M. C., & Hamilton, D. P. 2002, *MNRAS*, 330, 232
- Moody, K., & Sigurdsson, S. 2009, *ApJ*, 690, 1370
- Noyola, E., Gebhardt, K., & Bergmann, M. 2008, *ApJ*, 676, 1008
- Ostriker, J. P., Spitzer, L. J., & Chevalier, R. A. 1972, *Astrophys J Lett*, 176, L51+
- Paresce, F., de Marchi, G., & Jędrzejewski, R. 1995a, *Astrophys J Lett*, 442, L57
- Paresce, F., de Marchi, G., & Romaniello, M. 1995b, *ApJ*, 440, 216
- Piotto, G. 2009, in *IAU Symposium*, Vol. 258, ed. E. E. Mamajek, D. R. Soderblom, & R. F. G. Wyse, 233–244

- Piotto, G., De Angeli, F., King, I. R., et al. 2004, *Astrophys J Lett*, 604, L109
- Piotto, G., Villanova, S., Bedin, L. R., et al. 2005, *ApJ*, 621, 777
- Piotto, G., Bedin, L. R., Anderson, J., et al. 2007, *Astrophys J Lett*, 661, L53
- Plummer, H. C. 1911, *MNRAS*, 71, 460
- Pooley, D., Lewin, W. H. G., Anderson, S. F., et al. 2003, *Astrophys J Lett*, 591, L131
- Portegies Zwart, S. F., & McMillan, S. L. W. 2002, *ApJ*, 576, 899
- Portegies Zwart, S. F., Baumgardt, H., Hut, P., Makino, J., & McMillan, S. L. W. 2004, *Nature*, 428, 724
- Portegies Zwart, S., McMillan, S., & Gieles, M. 2010, *ArXiv e-prints*
- Press, W. H., & Teukolsky, S. A. 1977, *ApJ*, 213, 183
- Quinlan, G. D. 1996, *New Astron*, 1, 255
- Rasio, F. A., & Shapiro, S. L. 1991, *ApJ*, 377, 559
- Redmount, I. H., & Rees, M. J. 1989, *Comments Astrophys*, 14, 165
- Shara, M. M., Drissen, L., Bergeron, L. E., & Paresce, F. 1995, *ApJ*, 441, 617
- Sigurdsson, S., & Hernquist, L. 1993, *Nature*, 364, 423
- Sigurdsson, S., & Phinney, E. S. 1995, *ApJS*, 99, 609
- Sills, A., Lombardi, Jr., J. C., Baily, C. D., et al. 1997, *ApJ*, 487, 290
- Spitzer, L. J. 1969, *Astrophys J Lett*, 158, L139+
- Spitzer, L. 1987, *Dynamical evolution of globular clusters*, ed. L. Spitzer
- Ventura, P., D'Antona, F., Mazzitelli, I., & Gratton, R. 2001, *Astrophys J Lett*, 550, L65
- Verbunt, F., & Hut, P. 1987, in *IAU Symposium, Vol. 125, The Origin and Evolution of Neutron Stars*, ed. D. J. Helfand, & J.-H. Huang, 187-+
- Verbunt, F., & Meylan, G. 1988, *A&A*, 203, 297
- Vesperini, E. 1998, *MNRAS*, 299, 1019
- Vesperini, E., & Heggie, D. C. 1997, *MNRAS*, 289, 898
- Vesperini, E., McMillan, S. L. W., & Portegies Zwart, S. 2009, *ApJ*, 698, 615

18 Dynamics of Disks and Warps

J. A. Sellwood

Department of Physics and Astronomy, Rutgers, The State
University of New Jersey, Piscataway, NJ, USA

1	<i>Introduction</i>	925
2	<i>Preliminaries</i>	926
2.1	Relaxation Rate	926
2.2	Mathematical Formulation	927
2.3	Orbits	928
2.4	Resonances	929
2.5	Wave-Particle Scattering	930
2.6	Vertical Motion	930
3	<i>Local Theory of Density Waves</i>	931
3.1	Plane Waves in a Thin Mass Sheet	931
3.2	Dispersion Relations	931
3.3	Axisymmetric Stability	932
3.4	Self-consistent Density Waves	933
3.5	Group Velocity	934
3.6	Swing Amplification	935
4	<i>Bar Instability</i>	937
4.1	Mechanism for the Bar Mode	938
4.2	Predicted Stability	939
4.3	Residual Concerns	940
5	<i>Lop-sided Modes</i>	940
6	<i>Groove and Edge Modes</i>	941
7	<i>Spiral Structure</i>	944
7.1	Spirals as Global Modes of Smooth Disks	945
7.2	Recurrent Transients	946
7.3	Spirals as Responses to Density Fluctuations	946
7.4	Nonlinear Spiral Dynamics	947
7.5	Spirals in Global N -body Simulations	948
7.6	A Recurrent Instability Cycle?	949

8	<i>Buckling Instabilities and Warps</i>	950
8.1	Local Theory of Bending Waves	951
8.2	Global Bending Modes	953
8.3	Simulations of Buckling Modes	953
8.4	Disks in Halos	955
8.5	Misaligned Infall	955
8.6	Warps Driven by Tides	956
9	<i>Bars</i>	957
9.1	Origin of Bars	958
9.2	Frequency of Bars	958
9.3	Structure of Bars	959
9.4	Gas Flow	962
9.5	Bar Pattern Speeds	963
9.6	Bars Within Bars	964
9.7	Buckling of Bars	965
9.8	Dynamical Friction on Bars	966
9.8.1	Theory	966
9.8.2	Halo Density Constraint	968
9.8.3	Halo Density Reduction by Bars	968
10	<i>Secular Evolution Within Disks</i>	968
10.1	Heating by Spirals	970
10.2	Churning by Transient Spirals	971
10.3	Cloud Scattering	971
10.4	Black Holes in the Halo	973
10.5	Discussion	973
11	<i>Fragility of Disks</i>	974
12	<i>Conclusions</i>	975
13	<i>Appendix</i>	976
A13.1	Relaxation Time in Spheroids and Disks	976
	<i>Acknowledgments</i>	978
	<i>References</i>	978

Abstract: This chapter reviews theoretical work on the stellar dynamics of galaxy disks. All the known collective global instabilities are identified and their mechanisms described in terms of local wave mechanics. A detailed discussion of warps and other bending waves is also given. The structure of bars in galaxies, and their effect on galaxy evolution, is now reasonably well understood, but there is still no convincing explanation for their origin and frequency. Spiral patterns have long presented a special challenge, and ideas and recent developments are reviewed. Other topics include scattering of disk stars and the survival of thin disks.

Keywords: Galaxies: evolution, Galaxies: halos, Galaxies: spiral, Galaxies: kinematics and dynamics

1 Introduction

A significant fraction of stars in the universe reside in the rotationally supported disks of galaxies. Disks are mostly thin and flat, but they are often warped away from their principal plane in the outer parts. Disk galaxies usually manifest spiral patterns, and rather more than half host bars. Most, but not all, disk galaxies have a central bulge, perhaps also a thick stellar disk, and generally a small fraction of stars reside in a quasi-spherical stellar halo, while the central attraction at large distances from the center is dominated by a dark halo. The material in most disks overwhelmingly orbits in a single sense, although a small fraction of galaxies have been found to host substantial counter-rotating components.

This chapter is primarily concerned with the dynamics of rotationally supported disks of stars. Stellar disks are believed to have formed over time from gas that had previously settled into centrifugal balance in the gravitational well of the galaxy, and the process of star formation continues to the present day in most disk galaxies. While stars are the dominant dynamical component today, the small gas fraction (usually $\lesssim 10\%$ by mass) can still play an important dynamical role in some contexts.

Disk dynamics is a rich topic for two principal reasons: (a) the organized orbital motion facilitates gravitationally driven collective behavior and (b) outward transfer of angular momentum extracts energy from the potential well. Space limitations preclude a detailed development and this review will mostly be confined to a summary of the principal results and open issues. The derivations of the principal formulae can be found in the excellent textbook by Binney and Tremaine (2008, hereafter BT08). Furthermore, no attempt is made to cite every paper that relates to a topic.

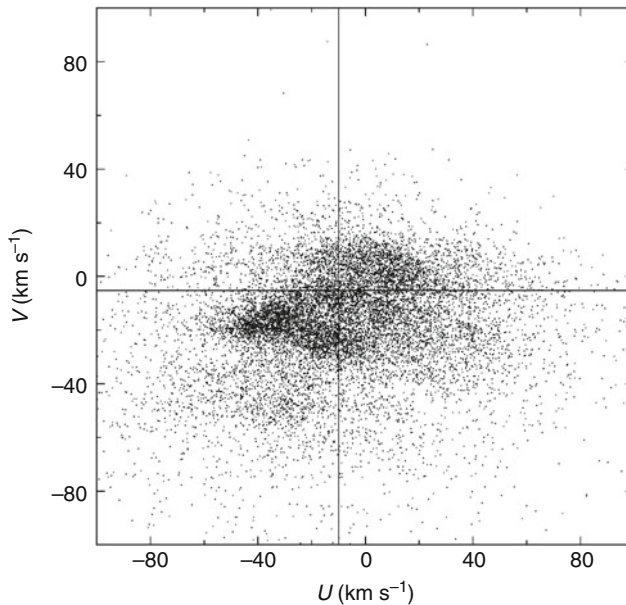
The distribution of gas within the Milky Way, the properties of the Galactic bulge, and other closely related topics are described elsewhere in this volume. The distributions of light and mass within galaxies and our current understanding of the processes that lead to the formation of galaxies are described in volume 6.

2 Preliminaries

2.1 Relaxation Rate

Because stellar disks contain many stars, the attraction from individual nearby stars is negligible in comparison with the aggregated gravitational field of distant matter. The Appendix explains how the usual rough calculation to support this assertion, which was derived with quasi-spherical systems supported by random motion in mind (e.g., BT08 pp. 35–38), must be revised for disks. These factors all conspire to reduce the relaxation time in disks by several orders of magnitude below the traditional estimate (☛ A4), although it remains many dynamical times.

☛ *Figure 18-1* shows the non-smooth distribution of stellar velocities of $>14,000$ F and G dwarf stars in the vicinity of the Sun, as found in the Geneva-Copenhagen Survey (Nordström et al. 2004; Holmberg et al. 2009 hereafter GCS). The determination of the radial velocities has confirmed the substructure that was first identified by Dehnen (1998) from a clever analysis of the HIPPARCOS data without the radial velocities (The implications of this figure are discussed more fully in ☛ Sect. 10.) As collisional relaxation would erase substructure, this distribution provides a direct illustration of the collisionless nature of the solar neighborhood (unless the substructure is being recreated rapidly, e.g., De Simone et al. 2004).



■ Fig. 18-1

The velocity distribution, in Galactic coordinates, of stars near the Sun, as given in the GCS. U is the velocity of the star toward the Galactic center and V is the component in the direction of Galactic rotation, both measured relative to the Sun. The intersection of the vertical and horizontal lines shows the local standard of rest estimated by Aumer and Binney (2009)

The usual first approximation that stars move in a smooth gravitational potential well therefore seems adequate.

2.2 Mathematical Formulation

This approximation immediately removes the need to distinguish stars by their masses. A stellar system can therefore be described by a **distribution function** (DF), $f(\mathbf{x}, \mathbf{v}, t)$ that specifies the stellar density in a 6D phase space of position \mathbf{x} and velocity \mathbf{v} at a particular time t . Since masses are unimportant, it is simplest conceptually to think of the stars being broken into infinitesimal fragments so that discreteness is never an issue.

With this definition, the mass density at any point is

$$\rho(\mathbf{x}, t) = \int f d^3\mathbf{v}, \quad (18.1)$$

which in turn is related to the gravitational potential, Φ , through **Poisson's equation**

$$\nabla^2\Phi(\mathbf{x}, t) = 4\pi G\rho(\mathbf{x}, t). \quad (18.2)$$

This is, of course, just the potential from the stellar component described in f ; the total potential includes contributions from dark matter, gas, external perturbations, etc. Finally, the evolution of the DF is governed by the **collisionless Boltzmann equation** (BT08 Eq. 4.11):

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{x}} + \dot{\mathbf{v}} \cdot \frac{\partial f}{\partial \mathbf{v}} = 0, \quad (18.3)$$

where the acceleration is the negative gradient of the smooth total potential: $\dot{\mathbf{v}} = -\nabla\Phi_{\text{tot}}$. The time evolution of a stellar system is completely described by the solution to these three coupled equations. Note that collisionless systems have no equation of state that relates the density to quantities such as pressure.

The most successful way to obtain global solutions to these coupled equations is through **N -body simulation**. The particles in a simulation are a representative sample of points in phase space whose motion is advanced in time in the gravitational field. At each step, the field is determined from a smoothed estimate of the density distribution derived from the instantaneous positions of the particles themselves.¹ This rough and ready approach is powerful, but simulations have limitations caused by **noise** from the finite number of particles, **bias** caused by the smoothed density and approximate solution for the field, and other possible artifacts.

Understanding the results from simulations, or even knowing when they can be trusted, requires dynamical insight that can be obtained only from analytic treatments. This chapter therefore stresses how the basic theory of stellar disks interrelates with well-designed, idealized simulations to advance our understanding of these complex systems.

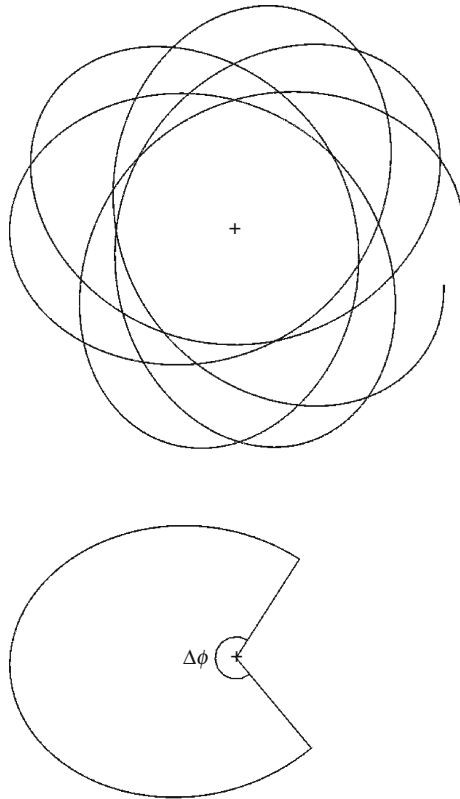
¹That is, a simulation solves (18.1–18.3) by the method of characteristics.

2.3 Orbits

When orbital deflections by mass clumps can be neglected (● Sect. 2.1), the stars in a disk move in a smooth gravitational potential. The effects of mass clumps and other non-uniformities are considered in ● Sect. 10. It is simplest to first discuss motion of stars in the disk mid-plane, before considering full 3D motion.

The orbit of a star of **specific energy** E and **specific angular momentum** L_z in the mid-plane of an axisymmetric potential is, in general, a non-closing rosette, as shown in ● Fig. 18-2. The motion can be viewed as a **retrograde epicycle** about a **guiding center** that itself moves at a constant rate around a circle of radius R_g , which is the radius of a circular orbit having the same L_z . In one complete radial period, τ_R , the star advances in azimuth through an angle $\Delta\phi$, as drawn in the lower panel. In fact, $\pi \leq \Delta\phi \leq 2\pi$ in most reasonable gravitational potentials; specifically, $\Delta\phi = \sqrt{2}\pi$ for small epicycles in a flat rotation curve.

These periods can be used to define two angular frequencies for the orbit: $\Omega_\phi = \Delta\phi/\tau_R$, which is the angular rate of motion of the guiding center, and $\Omega_R = 2\pi/\tau_R$. In the limit of the



■ Fig. 18-2

An orbit of a star is generally a non-closing rosette (upper panel). The lower panel shows one radial oscillation, from pericenter to pericenter say, during which time the star moves in azimuth through the angle $\Delta\phi$

radial oscillation amplitude, $a \rightarrow 0$, these frequencies tend to $\Omega_\phi \rightarrow \Omega$, the angular frequency of a circular orbit at $R = R_g$, and $\Omega_R \rightarrow \kappa$, the **epicyclic frequency**, defined through

$$\kappa^2(R_g) = \left(R \frac{d\Omega^2}{dR} + 4\Omega^2 \right)_{R_g}. \quad (18.4)$$

2.4 Resonances

If the potential includes an infinitesimal non-axisymmetric perturbation having m -fold rotational symmetry and which turns at the angular rate Ω_p , the **pattern speed**, stars in the disk encounter wave crests at the Doppler-shifted frequency $m|\Omega_p - \Omega_\phi|$. Resonances arise when

$$m(\Omega_p - \Omega_\phi) = l\Omega_R, \quad (18.5)$$

where simple orbit perturbation theory breaks down for steady potential perturbations. At **corotation** (CR), where $l = 0$, the guiding center of the star's orbit has the same angular frequency as the wave. At the **inner Lindblad resonance** (ILR), where $l = -1$, or at the **outer Lindblad resonance** (OLR), where $l = +1$, the guiding center respectively overtakes, or is overtaken by, the wave at the star's unforced radial frequency. Other resonances arise for larger $|l|$, but these three are the most important. Note that resonant orbits close after m radial oscillations and l turns about the galactic center in a frame that rotates at angular rate Ω_p .

The solid curve in **Fig. 18-3**, the so-called Lindblad diagram, marks the locus of circular orbits for stars in an axisymmetric disk having a flat rotation curve. Stars having more energy

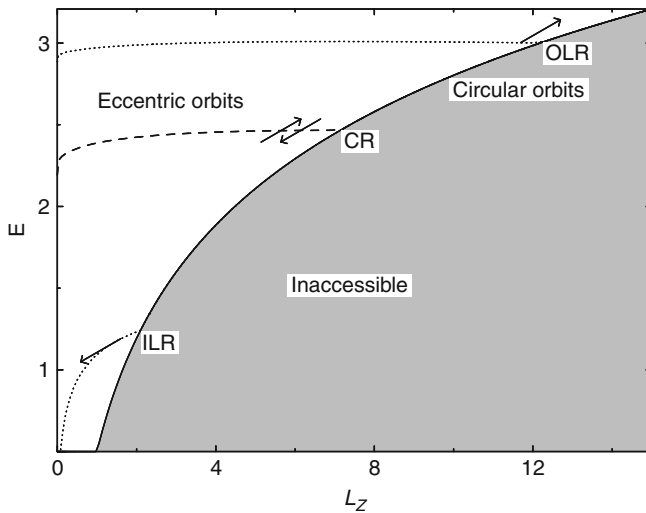


Fig. 18-3

The Lindblad diagram for a disk galaxy model. Circular orbits lie along the full-drawn curve and eccentric orbits fill the region above it. Angular momentum and energy exchanges between a steadily rotating disturbance and particles move them along lines of slope Ω_p as shown. The *dotted* and *dashed lines* are the loci of resonances for an $m = 2$ perturbation of arbitrary pattern speed

for their angular momentum pursue noncircular orbits. The broken curves show the loci of the principal resonances for some bi-symmetric wave having an arbitrarily chosen pattern speed; thus resonant orbits, which close in the rotating frame, persist even to high eccentricities.

2.5 Wave-Particle Scattering

The arrows in [Fig. 18-3](#) indicate how these classical integrals are changed for stars that are scattered by a steadily rotating mild potential perturbation. Since **Jacobi's invariant**,

$$I_J \equiv E - \Omega_p L_z \quad (18.6)$$

(BT08 Eq. 3.112), is conserved in axes that rotate with the perturbation, the slope of all scattering vectors in [Fig. 18-3](#) is $\Delta E/\Delta L_z = \Omega_p$.

Lynden-Bell and Kalnajs (1972) showed that a steadily rotating potential perturbation causes lasting changes to the integrals only at resonances. Since $\Omega_c = \Omega_p$ at CR, scattering vectors are parallel to the circular orbit curve at this resonance, where angular momentum changes do not alter the energy of random motion. Outward transfer of angular momentum involving exchanges at the Lindblad resonances, on the other hand, does move stars away from the circular orbit curve, and enables energy to be extracted from the potential well and converted to increased random motion.

Notice also from [Fig. 18-3](#) that the direction of the scattering vector closely follows the resonant locus (dotted curve) at the ILR only. Thus, when stars are scattered at this resonance, they stay on resonance as they gain random energy, allowing strong scatterings to occur. The opposite case arises at the OLR, where a small gain of angular momentum moves a star off resonance.

2.6 Vertical Motion

Stars also oscillate vertically about the disk mid-plane. If the vertical excursions are not large, and the in-plane motion is nearly circular, the vertical motion is harmonic with the characteristic frequency $\nu = (\partial^2 \Phi / \partial z^2)^{1/2}$, which principally depends on the density of matter in the disk mid-plane at R_g . As ν is significantly larger than κ , the frequency of radial oscillation, the two motions can be considered to be decoupled for small-amplitude excursions in both coordinates. In practice, the vertical excursions of most stars are large enough to take them to distances from the mid-plane where the potential departs significantly from harmonic (i.e., for $|z| \gtrsim 200$ pc in the solar neighborhood, Holmberg and Flynn 2004).

The principal **vertical resonances** with a rotating density wave occur where $m(\Omega_p - \Omega_\phi) = \pm \nu$, which are at radii farther from corotation than are the in-plane Lindblad resonances. Since spiral density waves are believed to have substantial amplitude only between the Lindblad resonances, the vertical resonances do not affect density waves, and anyway occur where the forcing amplitude will be small.

The situation is more complicated when the star's radial amplitude is large enough that ν varies significantly between peri- and apo-galacticon. Also, the vertical periods of stars lengthen for larger vertical excursions, allowing vertical resonances to become important (see [Sect. 8.1](#)).

3 Local Theory of Density Waves

3.1 Plane Waves in a Thin Mass Sheet

Since solutions of Poisson's equation can be obtained analytically only for the simplest mass distributions, theoretical treatments necessarily make quite drastic simplifying assumptions. Local theory, summarized in this section, is built around a WKB-type potential solution for density variations in the form of short-wavelength plane waves in a locally uniform, razor thin disk. In this case, a plane-wave disturbance $\Sigma_1 e^{ikx}$ to the surface density in the (x, y) -plane at some instant gives rise to the potential

$$\Phi_1(\mathbf{x}) = -\frac{2\pi G}{|k|} \Sigma_1 e^{ikx} e^{-|kz|}, \quad (18.7)$$

(cf. Eq. 6.31 of BT08). This relation strictly applies only to straight waves of infinite extent. However, the sinusoidal density variations cause the field of distant waves to cancel quickly, and the formula is reasonably accurate near the center of a short-wave packet. Note that (18.7) does not depend on the inclination of the wavefronts to the radial direction, but the curvature of the wavefronts must also be negligible, which requires $|kR| \gg 1$, with R being the distance from the disk center.

3.2 Dispersion Relations

A dispersion relation gives the relationship between the wavenumber k and frequency ω of self-consistent waves. Assuming (a) the above relation between density and potential for a razor-thin disk and (b) the wave is of small amplitude, so that second-order terms are negligible, the local dispersion relation for short axisymmetric density waves may be written as

$$\omega^2 = \kappa^2 - 2\pi G \Sigma |k| \mathcal{F}(s, \chi) \quad (18.8)$$

Kalnajs (1965), where Σ is the local undisturbed surface density and the complicated **reduction factor** \mathcal{F} , with two dimensionless arguments, is explained below. Lin and Shu (1966) independently derived a generalized relation for *tightly-wrapped* spiral waves, which required two additional assumptions: (c) that the wave vector k is approximately radial and (d) that the disturbance is not close to any of the principal resonances. Their better-known relation simply replaces the frequency of the wave with the Doppler-shifted forcing frequency experienced by the orbit guiding center:

$$(\omega - m\Omega)^2 = \kappa^2 - 2\pi G \Sigma |k| \mathcal{F}(s, \chi) \quad (18.9)$$

(see also BT08 Sect. 6.2.2), where now $\omega \equiv m\Omega_p$.

The reduction factor, \mathcal{F} , is always positive and is unity when the disk has no random motion. Since all the factors in the self-gravity term on the RHS are intrinsically positive, (18.8) and (18.9) say that self-gravity enables a supporting response from the stars at a frequency that is lower than their free epicycle frequency κ .

When the disk has random motion, the vigor of the stellar response depends on two factors: the ratio of the forcing frequency experienced by a star to its natural frequency, $s \equiv |\omega - m\Omega|/\kappa$, and the ratio of the typical sizes of the stellar epicycles ($\propto \langle v_R^2 \rangle^{1/2}/\kappa$) to the wavelength of the

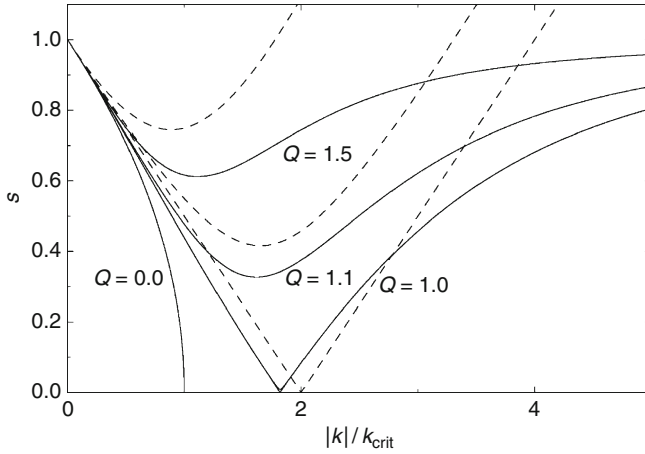


Fig. 18-4

The *solid lines* show the dispersion relation (18.9), while the *dashed lines* are from (18.10), each for several values of Q . The ordinate $s = |\omega - m\Omega|/\kappa$. Since the signs of both the abscissa and the ordinate are suppressed, the graph is reflection symmetric about both axes. Note that $s = 0$ at CR and $s = 1$ at the Lindblad resonances

wave ($\propto |k|^{-1}$). Thus for a Gaussian distribution of radial velocities, \mathcal{F} is a function of both s and $\chi \equiv \sigma_R^2 k^2 / \kappa^2$. Clearly, when χ is large, the unforced epicyclic amplitude of most stars is larger than the radial wavelength, and the weakened supporting response arises mainly from the small fraction of stars near the center of the velocity distribution.

The dispersion relation for barotropic fluid (gaseous) disks² is similar (BT08 Eq. 6.55):

$$(\omega - m\Omega)^2 = \kappa^2 - 2\pi G\Sigma|k| + v_s^2 k^2, \tag{18.10}$$

where v_s is the sound speed in the gas. Both dispersion relations are shown graphically in Fig. 18-4.

3.3 Axisymmetric Stability

Toomre (1964) determined the condition for marginal stability of short axisymmetric waves by solving for $\omega^2 = 0$ in (18.8). Since all the factors in the self-gravity term are intrinsically positive, ω^2 could be negative for large $|k|$, implying instability. Without random motion, $\mathcal{F} = 1$ and short waves with $k > k_{\text{crit}}$ or $\lambda < \lambda_{\text{crit}} = 4\pi^2 G\Sigma/\kappa^2$, will be unstable. Unlike for the Jeans instability in a stationary medium, longer waves are stabilized by rotation, embodied in the κ^2 term.

For disks with random motion, the reduction factor $\mathcal{F} \rightarrow 0$ for large $|k|$, irrespective of the frequency, for the reason given above. Thus random motion stabilizes short waves, as for

²Goldreich and Lynden-Bell (1965a) derived the vertically integrated equations for “2D pressure” in a gas disk.

the Jeans instability. Toomre (1964) showed for a Gaussian distribution of radial velocities that the RHS of (18.8) is ≥ 0 for all $|k|$ when $\sigma_R \geq \sigma_{R,\text{crit}}$, where

$$\sigma_{R,\text{crit}} \simeq \frac{3.36G\Sigma}{\kappa}. \quad (18.11)$$

Thus **Toomre's local axisymmetric stability criterion** is $Q \equiv \sigma_R/\sigma_{R,\text{crit}} \geq 1$. When reasonably constant with radius, the locally estimated Q value is a good indicator of global axisymmetric stability. For example, the local values in the equilibrium models proposed by Kalnajs (1976) are in reasonable agreement with those he derived from global axisymmetric modes (his Fig. 2). It should be noted that an axisymmetrically stable disk, i.e., for which $Q \geq 1$ everywhere, could still be unstable to non-axisymmetric modes.

The local axisymmetric stability criterion for rotating gas disks is similar. The longest unstable wavelength, λ_{crit} in a cold ($v_s = 0$) disk is the same as for a stellar disk. The minimum of the quadratic expression in k on the RHS of (18.10) occurs for $k = \pi G\Sigma/v_s^2$, and the condition of marginal stability ($\omega^2 = 0$) at this minimum is readily solved, yielding $v_{s,\text{crit}} = \pi G\Sigma/\kappa$.

Vandervoort (1970), Romeo (1992), and others have extended the discussion of axisymmetric stability to disks of finite thickness. The principal difference is the dilution of the self-gravity term caused by spreading the disk matter in the vertical direction, resulting in a somewhat more stable disk. The precise correction depends on the assumed vertical structure of both the disk and the wave, but is minor when the characteristic disk thickness $z_0 \ll \lambda_{\text{crit}}$, which is usually the case.

More realistic composite disks consist of multiple stellar populations having a range of velocity dispersions, as well as a cool gas component. Jog and Solomon (1992) considered the stability of two gravitationally coupled gas disks having different sound speeds, but Rafikov (2001) improved their analysis to include multiple stellar components. These analyses show that while the gas component may contain a small fraction (typically $\lesssim 10\%$) of the total mass, its low sound speed causes it to have a disproportionate effect on the overall stability.

3.4 Self-consistent Density Waves

Figure 18-4 gives, for four different values of Q , a graphical representation of the stellar (solid curves)³ and gaseous (dashed curves) dispersion relations (18.9) and (18.10), respectively. The curves do not differ for $Q = 0$ for which s is imaginary when $k > k_{\text{crit}}$. The dashed and solid curves have similar forms for moderate k but differ substantially for large k , where frequencies in the stellar sheet cannot exceed $|\omega - m\Omega| = \kappa$. Thus waves in a stellar sheet extend only from CR, where $s = 0$, to the Lindblad resonance on either side, where $s = 1$.⁴

A gas layer, on the other hand, can support gravitationally modified sound waves of arbitrarily high frequency. Note that this important difference in the stellar case can lead to artificial modes when hydrodynamic equations are used as a simplifying approximation for a stellar disk; while meaningful results can be obtained using this approximation, it is important to check that the derived modes do not cross Lindblad resonances.

³The reduction factor used assumes a Gaussian distribution of radial velocities.

⁴Additional Bernstein-type waves exist near integer values of $s > 1$, but such solutions seem to be of little dynamical importance.

Since the dispersion relation (● 18.9) does not depend on the sign of k , curves for negative k , or leading waves, are simple reflections about the $k = 0$ axis. Furthermore, the relation gives the square of $(\omega - m\Omega)$, implying that solutions when this frequency is negative, which arise inside CR, are simple reflections about the $s = 0$ axis. Thus were the signs not suppressed, ● Fig. 18-4 shows only the panel for trailing waves outside CR, while the other quadrants would be reflected images.

In the quadrant shown, there are either two values of k for each value of s , known as the **short-** and **long-wave branches**, or there are none. Only for the marginally stable case of $Q = 1$ do solutions exist for all frequencies $0 \leq s \leq 1$ and all $|k|$. When $Q > 1$, a **forbidden region** opens up in the vicinity of CR where waves are evanescent.

Recall that the relation (● 18.9) was derived assuming both $|kR| \gg 1$ and that the waves are tightly wrapped – i.e., that the wave vector is directed radially. These approximations may not be too bad on the short-wave branch (large k) but long-wave branch solutions may not exist at all, except in very low-mass disks where k_{crit} is large, and *must* fail as $|k| \rightarrow 0$. Lynden-Bell and Kalnajs (1972) gave an improved dispersion relation (their Eq. A11) that relaxed the requirement that the wave vector be radial, but still used the plane-wave potential (● 18.7) that requires $|kR| \gg 1$. Their relation is no different for tightly wrapped waves, but has no long-wave branch or forbidden region for open waves when $Q \gtrsim 1$.

3.5 Group Velocity

The waves in a disk described by the local dispersion relations (● 18.9) and (● 18.10) have a phase speed, equal to the pattern speed, in the azimuthal direction. However, Toomre (1969) pointed out that these dispersive waves also have a group velocity that is directed radially. Since $v_g = \partial\omega/\partial k$, the group velocity is proportional to the slope of the lines in ● Fig. 18-4. Each portion of a wavetrain, or wave packet, is carried radially at the group velocity, retaining its pattern speed ω while gradually adjusting its wavelength, and also transporting **wave action** or angular momentum.

For trailing waves outside CR, the situation illustrated in ● Fig. 18-4, the group velocity is positive on the short-wave branch, and the waves carry angular momentum outward toward the OLR. The curves have the opposite slope inside CR, where density waves are disturbances of negative angular momentum (stars there lose angular momentum as the wave is set up, Lynden-Bell and Kalnajs 1972), and therefore the inward group velocity on the short-wave branch leads once more to outward angular momentum transport. Outward angular momentum transport is in the sense expected from the gravitational stresses of a trailing spiral (Lynden-Bell and Kalnajs 1972).

However, the sign of the slope $\partial\omega/\partial k$ on the dubious long-wave branch is opposite to that of short waves, indicating that angular momentum in that regime is transported in the direction opposite to the gravity torque! This apparent contradiction is resolved by the phenomenon of **lorry transport**, a term coined by Lynden-Bell and Kalnajs (1972) to describe an advective Reynolds stress (see also Appendix J of BT08). Stars in their epicyclic oscillations gain angular momentum from the wave near apo-center and lose it back to the wave near peri-center; no star suffers a net change, but angular momentum is still carried inward at a sufficient rate to overwhelm the gravity torque.

Thus, where (● 18.9) holds in a disk with approximately uniform $Q \gtrsim 1$, a *tightly wrapped* trailing wave packet originating on the long-wave branch will travel towards CR until it reaches the edge of the forbidden zone where it “refracts” into a short wave that carries it back toward

the Lindblad resonance. The details of the turning point at the forbidden zone, which requires an evanescent wave propagating into the forbidden region, were described by Mark (1976) and by Goldreich and Tremaine (1978).

The fate of the short-wave propagating toward the Lindblad resonance also requires more sophisticated analysis because the dispersion relation (18.9) does not hold near resonances. Mark (1974) carried the analysis to second order to show that the wave is absorbed there through the wave-particle coupling described by Lynden-Bell and Kalnajs (1972).

3.6 Swing Amplification

Goldreich and Lynden-Bell (1965b), Julian and Toomre (1966) and Toomre (1981), extended the above analysis for tightly wrapped waves to waves of arbitrary inclination. They again adopted the approximate potential of a short-wavelength plane wave (18.7) and focused on a local patch of the disk whose dimensions were assumed small compared with R_0 , the radial distance of the patch center, which orbits the disk center at the local circular speed $V = R_0\Omega_0$.

Following Hill (1878), they then introduced Cartesian-like coordinates, $x \equiv R - R_0$ and $y \equiv R_0(\phi - \phi_0)$, where ϕ_0 is the azimuth of the orbiting center of the patch. The vertical coordinate, z , is unchanged from the usual cylindrical polar coordinate. Since the angular rotation rate decreases outward in galaxies, the motion of stars in the patch is that of a shear flow, with a speed that varies as $\dot{y} = -2Ax$, together with Coriolis forces arising from the rotation of the patch at angular rate Ω_0 . The Oort “constant” $A = -\frac{1}{2}Rd\Omega/dR$. The undisturbed surface density of matter, Σ , is assumed constant over the patch. This set of approximations is described as the **sheared sheet** (see BT08 pp. 678–681).

If α is the inclination angle of a plane density wave to the azimuthal direction, with $180^\circ > \alpha > 90^\circ$ for leading waves and $90^\circ > \alpha > 0^\circ$ for trailing waves, then the shear causes the pitch angle to change with time as $\cot \alpha = 2At$, with $t = 0$ when $\alpha = 90^\circ$, i.e., when the wave is purely radial. Both Goldreich and Lynden-Bell (1965b) for a gaseous sheet and Julian and Toomre (1966) for a stellar sheet found waves amplify strongly, but transiently, as they shear from leading to trailing.

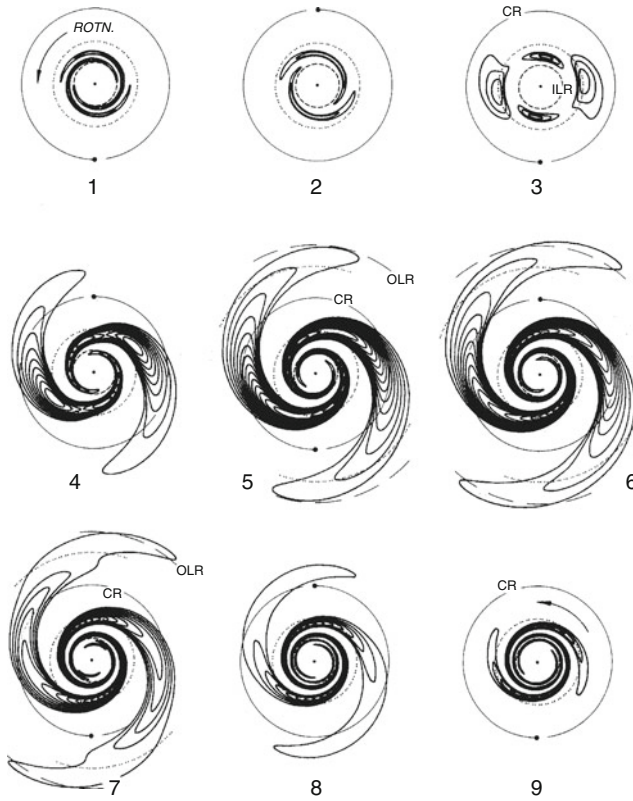
The unremitting shear flow ultimately tears apart any transient disturbance, and indeed Julian and Toomre (1966) asserted that the stellar sheet is locally stable at small amplitudes when $Q \geq 1$, but they did not give the proof.

Figure 18-5, from Toomre (1981), illustrates the dramatic transient trailing spiral that results from a small input leading disturbance. This calculation illustrates that even a global, small amplitude disturbance does not persist, but decays after its transient flourish. In the late stages, the “wave action” propagates at the group velocity (Sect. 3.5) away from corotation, where it is absorbed at the Lindblad resonances.

Notice also that the wave in this $Q = 1.5$ disk extends across the forbidden region near CR. The absence of solutions near $s = 0$ when $Q > 1$ in Fig. 18-4 is a consequence of assuming a steady, tightly wrapped wave; shearing waves do not respect this restriction.

The responsiveness of a disk to input leading signal depends both on the degree of random motion – it decreases rapidly over the range $1 < Q < 2$ – and upon the azimuthal wavelength, which is $\lambda_y = 2\pi R/m$ for an m -armed pattern. Julian and Toomre (1966) defined the parameter

$$X \equiv \frac{\lambda_y}{\lambda_{\text{crit}}} = \frac{Rk_{\text{crit}}}{m}. \quad (18.12)$$



■ Fig. 18-5

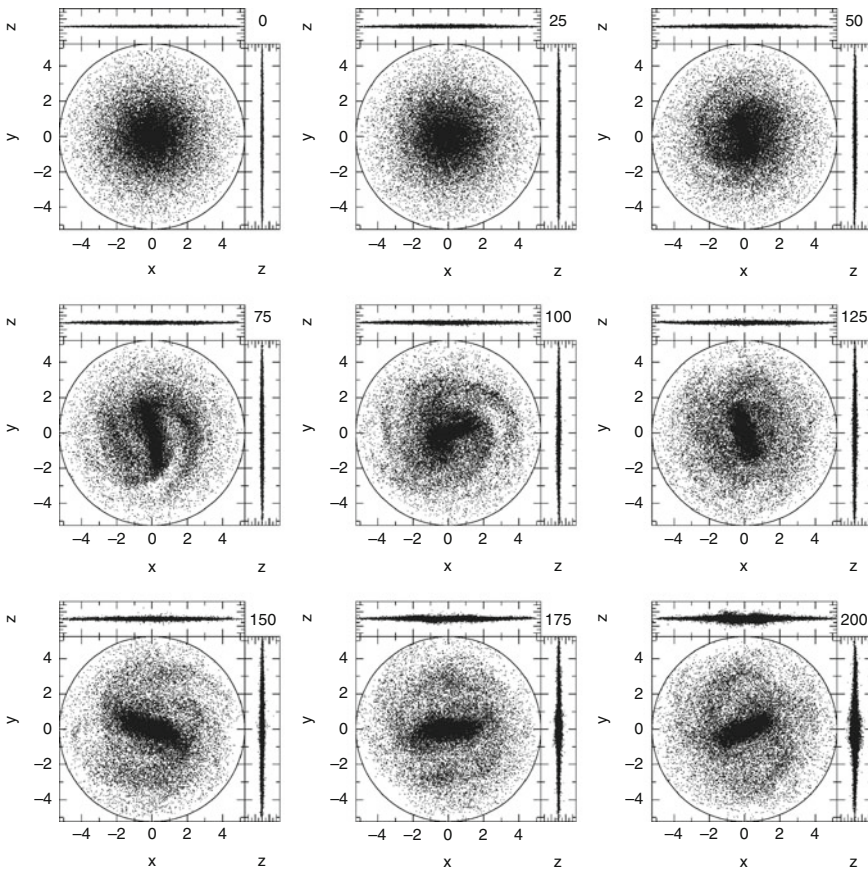
The time evolution of an input leading wave packet in the half-mass Mestel disk. The unit of time is half a circular orbit period at the radius marked CR. Reproduced from Toomre (1981)

In a disk having a flat rotation curve, Toomre (1981) showed that the strongest amplification occurs when $1 \lesssim X \lesssim 2.5$, or about *twice* the scale of the longest unstable axisymmetric wave. The peak amplification shifts to larger X values for declining rotation curves and vice versa for rising ones. Naturally, the phenomenon of swing amplification disappears entirely in uniformly rotating disks.

By making the assumption that spiral patterns in galaxies are amplified solely by this mechanism, Athanassoula et al. (1987) argued that the multiplicity of the observed spiral arm pattern can be used to place bounds on the mass density in the disk. The circular orbit speed in a nearly axisymmetric galaxy affords a direct estimate of the central attraction, but it is hard to determine how much of the total attraction is contributed by the disk, and how much by other components, such as the dark matter halo. Because λ_{crit} is proportional to the disk surface density, for a given rotation curve, the most vigorous amplification shifts to higher sectoral harmonics in lower mass disks. Generally, disks that contribute most of the central attraction, i.e., close to **maximum disks**, would prefer bi-symmetric spirals, while higher arm multiplicities would indicate significantly submaximal disks (see also Sellwood and Carlberg 1984).

4 Bar Instability

The puzzle of how galaxy disks could be stable presented a major obstacle to the development of our understanding of disk dynamics for many years. Superficially reasonable models of disk galaxies were found repeatedly, both in simulations (e.g., Miller et al. 1970; Hohl 1971; Zang and Hohl 1978; Combes and Sanders 1981; Sellwood 1981; Athanassoula and Sellwood 1986; Khoperskov et al. 2007; Dubinski et al. 2009) and in linear stability analyses (Kalnajs 1978; Aoki et al. 1979; Toomre 1981; Sawamura 1988; Vauterin and Dejonghe 1996; Pichon and Cannon 1997; Korchagin et al. 2005; Polyachenko 2005; Jalali 2007), to possess vigorous, global and disruptive bar-forming instabilities. ▶ *Figure 18-6* illustrates the global, disruptive nature of this instability. While it is premature to claim that this problem has been completely solved, it now seems that the stability of a massive disk galaxy requires only a dense bulge-like mass




■ Fig. 18-6

The formation of a bar in an (unpublished) *N*-body simulation. The three orthogonal projections show only particles in the initially exponential disk, halo particles are omitted. The disk, which started with $Q = 1.2$, has unit mass and unit scale length and $G = 1$. One orbit at $R = 2$ takes 15 time units, where the central attractions of the disk and halo are nearly equal


component near the center and owes little to the inner density of a dark matter halo. Galaxies lacking a central mass concentration, however, are still believed to require significant inner halo mass for global stability.

4.1 Mechanism for the Bar Mode

Toomre (1981) provided the most important step forward by elucidating the mechanism of the bar instability (see also BT08 Sect. 6.3). Linear bar-forming modes are standing waves in a cavity, akin to the familiar modes of organ pipes and guitar strings. Reflections in galaxies take place at the center and at the corotation radius, except that outgoing leading spiral waves incident on the corotation circle are super-reflected into amplified ingoing trailing waves (i.e., swing amplification), while also exciting an outgoing transmitted trailing wave. The feedback loop is closed by the ingoing trailing wave reflecting off the disk center into a leading wave, which propagates outward because the group velocity of leading waves has the opposite sign to that of the corresponding trailing wave. The amplitude of the continuous wave-train at any point in the loop rises exponentially, because the circuit includes positive feedback.

Toomre supported this explanation with linear stability studies of two disks. The Gaussian disk, which has a low central density, manifests a set of modes in which the more slowly growing, higher “overtones” display the kind of standing wave pattern to be expected from the superposition of ingoing trailing and outgoing leading waves. (The eigenfrequencies of these modes are shown in  Fig. 18-9.)

The other linear stability study he presented was for the inappropriately named “Mestel” disk,⁵ whose unusual stability properties were first described by Toomre’s student Zang (1976) in an unpublished thesis, and later by Evans and Read (1998). This disk has the scale-free surface density profile $\Sigma = V_c^2 / (2\pi GR)$, with V_c being the circular orbital speed that is independent of R . Zang had carried through a global, linear stability analysis of this disk, with random motions given by a smooth distribution function. In order to break the scale-free nature of the disk, Zang introduced a central cutout, and later an outer taper, in the active mass density, replacing the removed mass by rigid components in order that the central attraction remained unchanged at all radii. The dominant linear instabilities he derived for the tapered disk were confirmed in N -body simulations by Sellwood and Evans (2001).

Zang (1976) showed that a full-mass Mestel disk is stable to bi-symmetric modes, even if $Q < 1$, provided the tapers are gentle enough. Evans and Read (1998) extended this important result to other power-law disks, finding they have similar properties. Such disks are not globally stable, however, because they suffer from lop-sided instabilities (see  Sect. 5). By halving the active disk mass, with rigid mass preserving the overall central attraction, Toomre (1981) was well able to eliminate the lop-sided mode from the Mestel disk. In fact, he claimed that when $Q = 1.5$, the half-mass Mestel disk was globally stable, making it the only known model of a nonuniformly rotating disk that is stable to all small-amplitude perturbations.

Cutting the feedback loop really does stabilize a disk (see next section), which seems to confirm Toomre’s cavity mode mechanism. Despite this, Polyachenko (2004) pointed out that the strong emphasis on phenomena at corotation places the blame for the instability on a radius well outside where the resulting bar has its peak amplitude. He therefore proposed an alternative mechanism for the bar instability based upon orbit alignment (Lynden-Bell 1979,

⁵Mestel (1963) solved the far greater challenge of a disk of finite radius that has an exactly flat rotation curve.

see also [Sect. 9.1](#)). Even though the mechanism was originally envisaged as a slow trapping process, Polyachenko (2004) argued it may also operate on a dynamical timescale, and he devised (Polyachenko 2005) an approximate technique to compute global modes that embodied this idea. While his method should yield the same mode spectrum as other techniques, his alternative characterization of the eigenvalue problem may shed further light on the bar-forming mechanism.

4.2 Predicted Stability

The fact that small-amplitude, bi-symmetric instabilities are so easily avoided in the Mestel disk, together with his understanding of bar forming modes in other models, led Toomre (1981) to propose that stability to bar formation merely required the feedback cycle through the center to be cut. He clearly hoped that a dense bulge-like mass component, which would cause an ILR to exist for most reasonable pattern speeds, might alone be enough to stabilize a cool, massive disk.

Unfortunately, this prediction appeared to be contradicted almost immediately by the findings of Efstathiou et al. (1982), whose N -body simulations formed similar bars on short timescales irrespective of the density of the central bulge component! They seemed to confirm previous conclusions (Ostriker and Peebles 1973) that only significantly submaximal disks can avoid disruptive bar-forming instabilities.

However, Sellwood (1989a) found that Toomre's prediction does not apply to noisy simulations because of nonlinear orbit trapping. Simulations in which shot noise was suppressed by quiet start techniques did indeed manifest the tendency toward global stability as the bulge was made more dense, as Toomre's linear theory predicted. Shot noise from randomly distributed particles, on the other hand, produced a large enough amplitude collective response for nonlinearities to be important, and a noisy simulation of a linearly stable disk quickly formed a strong bar, consistent with the results reported by Efstathiou et al. (1982).

Since density variations in the distribution of randomly distributed particles are responsible for bar formation in this regime, the reduced shot noise level from larger numbers of particles must result in lower-amplitude responses that ultimately should avoid nonlinear trapping. The precise particle number required depends on the responsiveness of the disk, which is weakened by random motion, lower surface density, increased disk thickness, or gravity softening. Efstathiou et al. (1982) employed merely 20,000 particles, which were clearly inadequate to capture the linear behavior. However, robustly stable, massive disks have been simulated by Sellwood and Moore (1999) and Sellwood and Evans (2001) that employed only slightly larger particle numbers. Note that these latter models also benefitted from more careful setup procedures to create the initial equilibrium.

Thus the stabilizing mechanism proposed by Toomre (1981) does indeed work in simulations of high enough quality and presumably, therefore, also in real galaxy disks. Indeed, Barazza et al. (2008) found a decreased incidence of bars in galaxies having luminous bulges, and argued that their result supports Toomre's stabilizing mechanism.

Thus the absence of bars in a significant fraction of high-mass disk galaxies does not imply that the disk is submaximal. The old stability criteria proposed by Ostriker and Peebles (1973), Efstathiou et al. (1982), and Christodoulou et al. (1995) apply only to disks that lack dense centers; indeed Evans and Read (1998) showed explicitly that the power-law disks are clear counterexamples to the simple bi-symmetric stability criterion proposed by Ostriker and Peebles (1973).

4.3 Residual Concerns

While all this represents real progress, a few puzzles remain. The most insistent is the absence of large, strong bars in galaxies like M33, which has a gently rising rotation curve. Although many spiral arms can be counted in blue images (Sandage and Humphreys 1980), the near-IR view (Block et al. 2004) manifests an open two-arm spiral, suggesting that the disk cannot be far from maximal, and also reveals a mild bar near the center of the galaxy. Corbelli and Waltherbos (2008) also found kinematic evidence for a weak bar. Perhaps the stability of this galaxy can be explained by some slightly larger halo fraction, or perhaps the weak bar has some unexpected effect, but there is no convincing study to demonstrate that this galaxy, and others like it, can support a two-arm spiral without being disruptively unstable.

The second concern is that lop-sided instabilities appear in extended full-mass disks with flat or declining rotation curves, which is discussed next. A third concern, which is discussed in ▶ Sect. 9.2, is that the mechanism is unable to predict the presence or absence of a bar in a real galaxy.

5 Lop-sided Modes

Many galaxies have apparently lop-sided disks. The treatment here will not go into detail, since Jog and Combes (2009) have recently reviewed both the observational data and theory.

Both theoretical and simulation work on $m = 1$ distortions to an axisymmetric disk require special care, since the absence of rotational symmetry can lead to artifacts unless special attention is paid to linear momentum conservation. Rigid mass components present particular difficulties, since they should not be held fixed, and extensive mass components are unlikely to respond as rigid objects.

As noted above, Zang (1976) found that the dominant instability of the centrally cut out Mestel disk was not the usual bar instability, but a lop-sided mode, which persists in a full mass disk no matter how large a degree of random motion or gentle the cutouts. This surprising finding was confirmed and extended to general power-law disks by Evans and Read (1998). A lop-sided instability dominated simulations Sellwood (1985) of a model having some resemblance to Zang's, in that it had a dense massive bulge and no extended halo, while Saha et al. (2007) reported similar behavior in simulations of a bare exponential disk. Lovelace et al. (1999) found pervasive lop-sided instabilities near the disk center in a study of the collective modes of a set of mass rings.

Various mechanisms have been proposed to account for this instability. Baldwin et al. (1980) and Earn and Lynden-Bell (1996) explored the idea that long-lived lop-sidedness could be constructed from cooperative orbital responses of the disk stars, along the lines discussed for bars by Lynden-Bell (1979). Tremaine (2005) discussed a self-gravitating secular instability in near-Keplerian potentials.⁶ A more promising mechanism is a cavity mode, similar to that for the bar-forming instability (Dury et al. 2008): the mechanism again supposes feedback to the swing amplifier, which is still vigorous for $m = 1$ in a full-mass disk.

⁶The “sling amplification” mechanism proposed by Shu et al. (1990) applies only to gaseous accretion disks, since it relies on sound waves propagating outside the OLR.

Feedback through the center cannot be prevented by an ILR for $m = 1$ waves, since the resonance condition $\Omega_p = \Omega_\phi - \Omega_R$ (► 18.5) can be satisfied only for retrograde waves. But the lopsided mode can be stabilized by reducing the disk mass, which reduces the X parameter (► 18.12) until amplification dies for $m = 1$ (Toomre 1981). Sellwood and Evans (2001) showed that together with a moderate bulge, the dark matter required for a globally stable disk need not be much more than a constant density core to the minimum halo needed for a flat outer rotation curve.

A qualitatively different lop-sided instability is driven by counter-rotation. This second kind of $m = 1$ instability was first reported by Zang and Hohl (1978) in a series of N -body simulations designed to explore the suppression of the bar instability by reversing the angular momenta of a fraction of the stars; they found that a lop-sided instability was aggravated as more retrograde stars were included in their attempts to subdue the bar mode. Analyses of various disk models with retrograde stars (Araki 1987; Sawamura 1988; Dury et al. 2008) have revealed that the growth rates of lop-sided instabilities increase as the fraction of retrograde stars increases. Merritt and Stiavelli (1990) and Sellwood and Valluri (1997) found lop-sided instabilities in simulations of oblate spheroids with no net rotation. Their flatter models had velocity ellipsoids with a strong tangential bias, whereas Sellwood and Merritt (1994) found that disks with half the stars retrograde, together with moderate radial motion were surprisingly stable.

Weinberg (1994) pointed out that lop-sided distortions to near spherical systems can decay very slowly, leading to a protracted period of “sloshing.” This **seiche mode** in a halo, which decays particularly slowly in mildly concentrated spherical systems, could provoke lop-sidedness in an embedded disk (Kornreich et al. 2002; Ideta 2002).

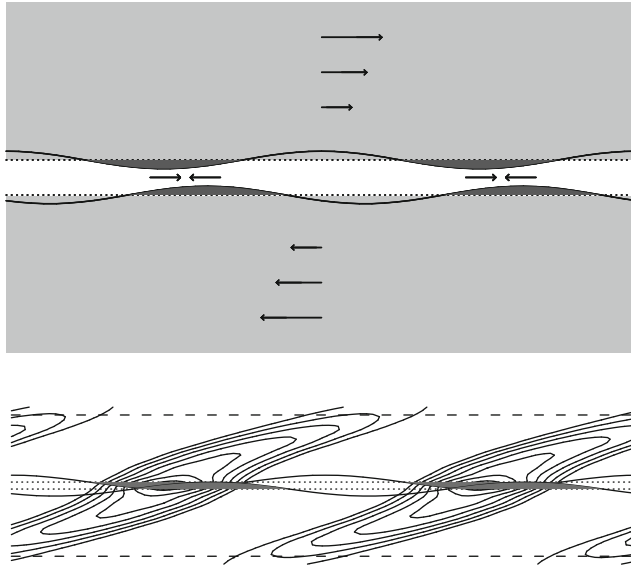
If lop-sidedness is due to instability, then the limited work so far suggests that it would imply a near-maximum disk. But lop-sidedness in the outer parts could also be caused by tidal interactions, or simply by asymmetric disk growth, with the effects of differential rotation being mitigated perhaps by the cooperative orbital responses discussed by Earn and Lynden-Bell (1996).

6 Groove and Edge Modes

Not all true global instabilities in disks require a feedback loop. Another class of modes is simply driven from corotation, with trailing waves propagating both inward and outward to be absorbed, in stellar disks, at the Lindblad resonances on either side.

Lovell and Hohl (1978) showed that disks are destabilized by a local extremum in the radial variation of the ratio of surface density to vorticity, $\Sigma/|\nabla \times \mathbf{v}| \equiv \Sigma\Omega/(2\kappa^2)$, or the reciprocal of potential vorticity, and proposed that the instability created flat rotation curves for which the potential vorticity is also flat. Sellwood and Kahn (1991) demonstrated that the instability caused by a rather insignificant, narrow, decrease in surface density, i.e., a “groove,” is a global spiral mode.

The mechanism is easiest to visualize in a disk without random motion, where small-scale surface density variations are not blurred by epicyclic motions. In this case, a deficiency of stars over a small range of angular momentum creates a groove in the surface density, as shown in the sheared sheet model in the top panel of ► Fig. 18-7. The groove itself is unstable because of the gravitational coupling between disturbances on either side. The dark shaded areas in the figure illustrate regions where small sinusoidal radial displacements of material on each edge have created high-density regions where the density was previously low. Disturbing gravitational



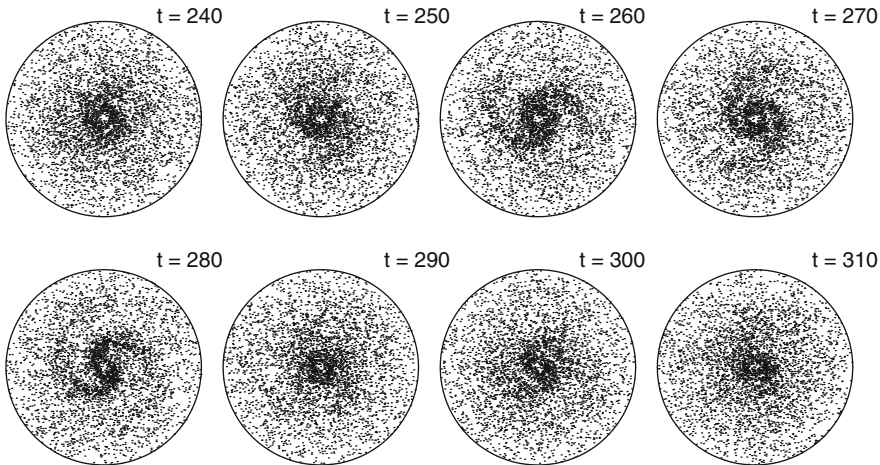
■ Fig. 18-7

The *top panel* illustrates a groove in the sheared sheet (▶ Sect. 3.6) model. The *light shaded* region has the full undisturbed disk surface density, Σ , which in this frame has a linearly varying shear flow. The *white* region is the groove, which has lower surface density, and the density excesses brought into the groove by small wave-like disturbances on either side of the groove are highlighted by the *dark shading*. The width of the groove is exaggerated for clarity. The *lower panel*, which has the correct aspect ratio, shows the supporting responses to the density variations on the groove edges. The *dashed lines* mark the locations of the Lindblad resonances

forces arise from the density excesses, as illustrated, which are directed along the groove if the wavelength is long compared with the groove width. Material that is pulled back loses angular momentum and sinks toward the center of the galaxy, while that which is urged forward gains and rises outward. Thus each density excess pulls on the other across the groove in such a way as to cause it to grow exponentially, i.e., the combined disturbance on the two sides is unstable. The groove edges need be only steep gradients, not discontinuities. Furthermore, the mechanism is the same, but harder to visualize, in a disk with random motion where the density of stars is depleted over a narrow range of angular momentum.

The growing disturbance in the groove creates wave-like mass variations along the groove that are effectively growing co-orbiting mass clumps of the type envisaged by Julian and Toomre (1966). Generalizing their apparatus to allow for exponentially growing masses, Sellwood and Kahn (1991) estimated the expected disk response, as shown in the lower panel of ▶ Fig. 18-7 for the parameters $Q = 1.8$ and $X = 2$, and low growth rate. The disk supporting response transforms the quite trivial disturbance in the groove into an extensive spiral instability!

Unlike the sheared sheet, the azimuthal wavelength in a full disk can take on only discrete values, $\lambda_y = 2\pi R/m$, which are all unstable, but the one which grows most rapidly is that for which swing amplification (▶ Sect. 3.6) is the most vigorous, i.e., $1 \lesssim X \lesssim 2.5$ (▶ 18.12) in a flat rotation curve. The simulation shown in ▶ Fig. 18-8 illustrates the scale and vigor of the mode, which saturates at $\sim 20\%$ overdensity because of the onset of horseshoe orbits at corotation (Sellwood and Binney 2002).



■ Fig. 18-8

The later part of the growth and subsequent decay of an isolated spiral mode in a disk that was seeded with a groove. Disk rotation is counterclockwise and disturbance forces in this simulation, taken from Sellwood and Binney (2002), were restricted to $m = 2$ only

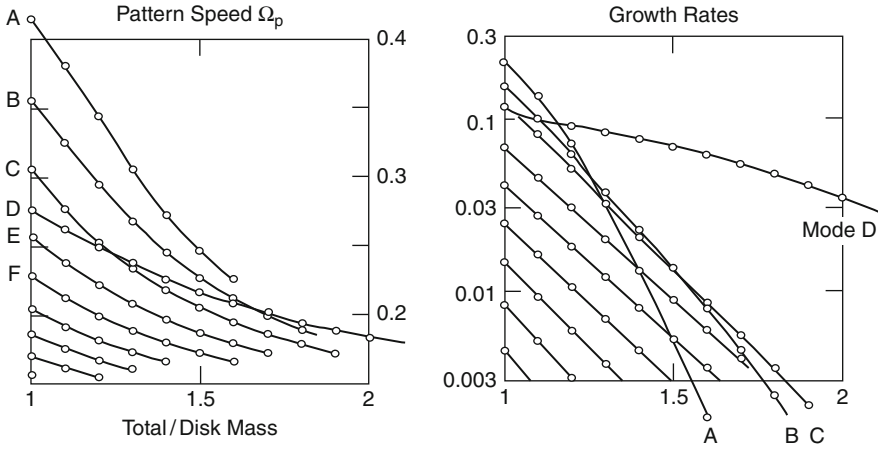
Lovelace and Hohlfield (1978) and Sellwood and Kahn (1991) found that almost any narrow feature in the angular momentum density is destabilizing, although the modes of a simple ridge, e.g., come in pairs with CR some distance from the ridge center.

Edge modes (Toomre 1981; Papaloizou and Lin 1989) are close cousins of groove modes, and the mechanism can be understood in the same fashion. Density variations on a single edge without a supporting disk response are neutrally stable, but the necessarily one-sided wakes in the massive interior disk add angular momentum to the density enhancements on the edge, causing them to grow.

The eigenfrequencies of the lower-order bi-symmetric modes of the Gaussian disk vary with decreasing disk mass as shown in Fig. 18-9, taken from Toomre (1981). Most are cavity modes, but the edge-mode, labeled “mode D” stands out because its growth rate in particular declines more slowly with decreasing disk mass. If the mechanism for both types of modes involves swing amplification, one might expect their vigor to be similarly affected by the rise in the X parameter as the surface density decreases. But the decreased surface density also slows the group velocity while the rapidly declining pattern speed of each cavity mode moves CR farther out in the disk. Thus the growth rates of cavity modes drop more quickly because of the increased travel time for the wave packet to complete the feed-back loop, a factor that is absent for the edge mode.

The edge instability requires only a steep gradient in the surface density, which need not drop to zero. Toomre (1989) gave the condition for instability as “the radial distance over which the disc density undergoes most of its rapid change should be no larger than about one quarter of the axisymmetric stability length λ_{crit} ,” assuming “the disc is massive and cool enough for vigorous swing-amplification.”

Curiously, global bar-forming modes were originally thought to be related to the rotational instabilities of uniform-density Maclaurin spheroids of incompressible fluid (Ostriker and Peebles 1973), an idea reinforced by the vigor of the bisymmetric mode of the sharp-edged



■ Fig. 18-9

The variation of pattern speeds and growth rates for the first few $m = 2$ modes of the Gaussian disk as the active disk mass is decreased. From Toomre (1981)

Maclaurin disk (Kalnajs 1972). Toomre (1981) noted that the instability of such unrealistic galaxy models may be more closely related to the edge mode than to the cavity mode described in Sect. 4.1.

As the consequence of an edge instability in a realistic galaxy model is to blur the edge, it seems unlikely that galaxy disks can retain unstable density gradients for interestingly long periods. Similarly, the large-amplitude evolution of groove or ridge modes quickly erases the feature that gave rise to them. Nevertheless, the modes have other possible consequences (see Sect. 7.6).

7 Spiral Structure

Despite many decades of effort, no theory to account for the graceful spiral patterns in disk galaxies is widely accepted. Most workers in this field agree that spiral patterns are gravitationally driven variations in surface density in the old stellar disk, as supported by photometric data (e.g., Grosbøl et al. 2004; Zibetti et al. 2009) and streaming motions in high spatial resolution velocity maps (e.g., Shetty et al. 2007).

There seems little doubt that some spiral patterns are tidally driven (e.g., Dobbs et al. 2010; Salo and Laurikainen 1993), while others could be the driven responses to bars (Buta et al. 2009). Although these two ideas may account for a large fraction of the cases (Kormendy and Norman 1979), especially if orbiting dark matter clumps can excite patterns (Dubinski et al. 2008), spirals can still develop in the absence of either trigger, as revealed in simulations.

The idea that spirals could be self-excited oscillations of the stellar disk represents the greatest theoretical challenge. While there is general agreement that gas seems to be essential, no picture seems complete, and current theories disagree on even the lifetimes of the patterns. One idea (e.g., Bertin and Lin 1996) is that spiral features are manifestations of quasi-steady

global modes of the underlying disk. Alternatively, they could be short-lived, recurrent transient patterns that originate either from forcing by density fluctuations (e.g., Toomre 1990) or else from recurrent vigorous instabilities (e.g., Sellwood 2000).

A serious obstacle to progress in this area has been the absence of observational discriminants that would favor one of these radically differing viewpoints over the other. The predictions for density variations or gas responses at a single instant are essentially independent of the generating mechanism and do not depend strongly on the lifetime of the pattern. Meidt et al. (2009) attempted to measure radial variations of pattern speeds using a generalization of the method devised by Tremaine and Weinberg (1984a) (see [Sect. 9.5](#)). They reported lower pattern speeds at larger radii, but their measurements still tell us little about spiral lifetimes. However, the velocity distribution of stars in the solar neighborhood (Nordström et al. 2004, and [Fig. 18-1](#)) is most naturally accounted for in the transient picture (see [Sect. 7.6](#)).

7.1 Spirals as Global Modes of Smooth Disks

Simple models of disk galaxies possess many linear instabilities (e.g., Jalali 2007). The bar-forming mode ([Sect. 4](#)) is usually the fastest growing, but it has almost no spirality. These studies are therefore important to understand stability, but even the higher modes are not particularly promising for spiral generation.⁷

The “density wave” theory for spiral modes, described in detail by Bertin and Lin (1996), invokes a more specific galaxy model with a submaximal disk that is dynamically cool in the outer parts and hot in the inner disk. The local stability parameter is postulated to be $1.0 \lesssim Q \lesssim 1.2$ over most of the disk and to rise steeply to $Q > 2$ near the center. Bertin, Lin, and their coworkers perform a global analysis using the hydrodynamic approximation (BT08 Sect. 5.1), which reveals slowly growing spiral modes under these specific conditions.

The mechanism (Mark 1977) is a cavity mode, having qualitative similarities to that for the bar mode ([Sect. 4.1](#)), but the tightly wrapped waves are trailing around the entire cycle. The inner turning point is at a Q -barrier: a steeply increasing Q value causes the forbidden zone ([Sect. 3.4](#)) to broaden to the point that ingoing short waves get “refracted” into outgoing long waves, which prevents the wave train from reaching the ILR where it would be damped. The long waves then propagate out to near the CR in the $Q \sim 1$ part of the disk, where they switch back to the short-wave branch with a small degree of amplification. The WASER mechanism (Mark 1976) at this turning point involves a third, transmitted wave that is “radiated” outward on the short-wave branch, carrying away the angular momentum to excite the mode in the inner disk. Thus the amplifier involves a small “swing” from the long- to the short-wave branch, whereas the bar instability uses a full swing from leading to trailing. This difference, together with their assumption that the disk is submaximal (i.e., $X > 3$ for $m = 2$, see [Sect. 18.12](#)), allows the mode to be slowly growing. Lowe et al. (1994) present a model of this kind to account for the spiral structure of M81.

⁷Korchagin et al. (2005) calculated essentially gas-dynamical modes for models of specific galaxies, and argued that the shapes of one, or more, of the lower-order modes could be matched to the observed spiral pattern. However, it is unclear that rapidly growing, linear modes can be seen for long at finite amplitude before nonlinear effects will change their appearance, and it seems even less likely that two modes with different growth rates should have similar large amplitudes at the time a galaxy is observed.

In order to justify the “basic state” of the disk they require, Bertin and Lin (1996) argued heuristically that rapidly evolving features would have disappeared long ago and that low-growth-rate instabilities in a cool disk, created by gas dissipative processes and star formation, will dominate at later times. They invoked shocks in the gas to limit the amplitude of the slowly growing mode, leading to a quasi-steady global spiral pattern.

The main objection to their picture is that it is likely that such a lively outer disk will suffer from other, more vigorous, collective responses with $m > 2$ that will quickly heat the disk, as described in [Sect. 7.5](#), and destroy their postulated background state.

7.2 Recurrent Transients

From the early work by Miller et al. (1970), Hohl (1971), Hockney and Brownrigg (1974), James and Sellwood (1978), and Sellwood and Carlberg (1984), N -body simulations of cool, submaximal disks have exhibited recurrent transient spiral activity. This basic result has not changed for several decades as numerical quality has improved.

Claims of long-lived spiral waves (e.g., Thomasson et al. 1990) have mostly been based on simulations of short duration. For example, Elmegreen and Thomasson (1993) presented a simulation that displayed spiral patterns for ~ 10 rotations, but the existence of some underlying long-lived wave is unclear because the pattern changed from snapshot to snapshot. Donner and Thomasson (1994) ran their simulations for fewer than seven disk rotations and the bi-symmetric spiral they observed appeared to be an incipient bar instability. Zhang (1996, 1998) adopted a similar mass distribution and also reported long-lived patterns in her simulations. The author has attempted to reproduce her results and indeed obtained similar bi-symmetric features, but they appear to be the superposition of several waves having differing pattern speeds.

Sellwood and Carlberg (1984) stressed that patterns fade in simple simulations that do not include the effects of gas dissipation; the reason is the disk becomes less responsive as random motion rises due to particle scattering by the spiral activity ([Sect. 10](#)), which is therefore self-limiting. They also demonstrated that mimicking the effects of dissipative infall of gas, such as by adding fresh particles on circular orbits, allowed patterns to recur “indefinitely.” Later work (e.g., Carlberg and Freedman 1985; Toomre 1990) has shown that almost any method of dynamical cooling can maintain spiral activity, as also happens in modern galaxy formation simulations (e.g., Agertz et al. 2010).

Thus the transient spiral picture offers a natural explanation for the absence of spiral patterns in $S0$ disk galaxies that have little or no gas; maintenance of spiral activity requires a constant supply of new stars on near-circular orbits. Other pieces of indirect evidence that also favor the transient spiral picture are reviewed in [Sect. 10](#).

7.3 Spirals as Responses to Density Fluctuations

Goldreich and Lynden-Bell (1965b) and Toomre (1990) suggested that a large part of the spiral activity observed in disk galaxies is the collective response of the disk to clumps in the density distribution. As a spiral wake is the collective response of a disk to an individual co-orbiting perturber (Julian and Toomre 1966), multiple perturbers will create multiple responses that all

orbit at different rates. The behavior of this polarized disk reveals a changing pattern of trailing spirals, which can equivalently be regarded as swing-amplified (Sect. 3.6) noise.⁸

Toomre and Kalnajs (1991) studied the amplified noise that arose in their N -body simulations of the sheared sheet. The massive particles themselves provoke spiral responses with an amplitude proportional to the input level of shot noise, caused by density variations in the distribution of randomly distributed particles. Comparison of their expectations with linear theory predictions revealed that the simulations were livelier than they expected, by a factor $\lesssim 2$, apparently from a gradual buildup of correlations between the mean orbital radii of the particles.

This could be a mechanism for chaotic spirals in gas-rich discs, where a high rate of dissipation may be able to maintain the responsiveness of the disk (Toomre 1990) while the clumpiness of the gas distribution may make the seed noise amplitude unusually high. However, it seems likely that spiral amplitudes (e.g., Zibetti et al. 2009) are too large to be accounted for by this mechanism in most galaxies. Also, the spiral structure should be chaotic, with little in the way of clear symmetry expected.

7.4 Nonlinear Spiral Dynamics

Tagger and his coworkers (Tagger et al. 1987; Sygnet et al. 1987; Masset and Tagger 1997) suggested that global modes in stellar disks could be coupled through nonlinear interactions. They proposed that wave 1 excites wave 2 through second-order coupling terms that are large when CR of wave 1 lies at approximately the same radius as the ILR of wave 2. Conservation rules require a third wave such that $m_3 = m_1 \pm m_2$, i.e., most likely an axisymmetric wave ($m_3 = 0$) if $m_1 = m_2$. Many examples of multiple waves in N -body simulations have been reported with remarkable coincidences for the radii of the main resonances.

Fuchs et al. (2005) developed a similar argument for waves in the sheared sheet. They found that amplified trailing waves could excite fresh leading waves in their second-order theory. They proposed this mechanism as an alternative source of the amplitude excess noted above in the simulations by Toomre and Kalnajs (1991), and they speculated that the same mechanism may also account for the larger than expected amplitudes of spirals in global N -body simulations (see Sect. 7.5). As both this mechanism and that discussed in the previous paragraph employ terms that are second-order in the perturbation amplitude, they become important only when features are strong.

Patsis et al. (1991) attempted to construct, by orbit superposition, self-consistent steady spiral waves of finite amplitude to match the observed non-axisymmetric patterns in specific galaxies. They experienced great difficulty in finding solutions near CR, and suggested that either this resonance or the 4:1 resonance⁹ marks the outer radius of the spiral. Their finding is not unexpected for two reasons: (a) the dispersion relation for steady, tightly wrapped,

⁸Cuzzi et al. (2010) found evidence for similar behavior within Saturn's A ring.

⁹The resonance condition (Sect. 18.5) for a pure $\cos(m\theta)$ potential variation also implies frequency commensurabilities at multiples of m . The small denominators that characterize the principal resonances (BT08, Sect. 3.3.3) arise at these "ultraharmonic resonances" only for noncircular orbits. A new family of orbits appears at the 4:1 resonance that closes after four radial oscillations (see Sect. 9.3 and Sellwood and Wilkinson 1993).

small-amplitude waves (► [Fig. 18-4](#)) predicts a forbidden region around CR for all $Q > 1$, and (b) the nonlinear dynamics of orbits in barred potentials (see ► [Sect. 9.3](#)) finds only chaotic orbits near CR, which are unfavorable to self-consistency as Patsis et al. (1991) found. It should be noted that the difficulty of finding a self-consistent solution near CR is a direct consequence of their assumption of a steady wave pattern; transient waves do not suffer from this problem (► [Fig. 18-5](#)) and rapidly growing groove modes (► [Sect. 6](#)) have *peak* overdensities at CR.

Tsoutsis et al. (2009) and Athanassoula et al. (2009) suggested that spirals in barred galaxies could be created by the slow migration of stars along Lyapunov manifold tube orbits emanating from the unstable Lagrange points at the ends of the bar.¹⁰ The ambitious hope of these preliminary papers is for an ultimate unified picture for the coexistence of bars, spirals, and rings, all having the same pattern speed.

7.5 Spirals in Global N -body Simulations

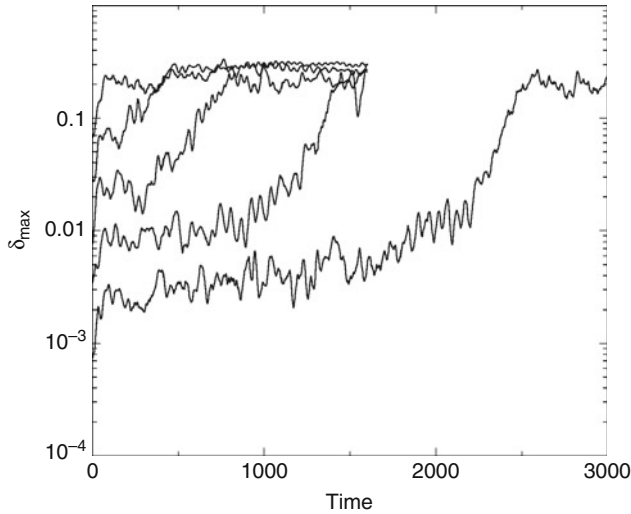
Sellwood and Carlberg (1984) and Sellwood (1989b) reported that their global simulations manifested more vigorous spiral activity than was consistent with amplified shot noise. A brief summary of some further results to support this claim is given here, and will be described more fully elsewhere.

As noted in ► [Sect. 4.1](#), Toomre (1981) predicted the half-mass Mestel disk to be globally stable to small-amplitude disturbances. Thus N -body simulations of this disk should exhibit no activity in excess of the inevitable swing-amplified shot noise. ► [Figure 18-10](#) reveals that this is not the case. The ordinate shows the largest value of $\delta = \Delta\Sigma/\Sigma$ from $m = 2$ disturbances in a sequence of simulations with increasing numbers of particles. The unit of time is R_i/V_c , where $V_c R_i$ is the center of the inner angular momentum cutout. Thus the orbit period at this small radius is 2π .

At $t = 0$, $\delta \propto N^{-1/2}$, as appropriate for shot noise, and swing amplification causes an almost immediate jump by a factor of a few for all N . When $N = 5 \times 10^4$, amplified noise causes ~20% overdensities almost immediately. For larger N the amplitude eventually rises to similar values in later evolution, once the inner disk has developed a pronounced bar. But for the largest two values of N shown, a period of slow growth occurs after the initial swing-amplified surge, offering tentative support for the linear theory prediction of global stability, with the slow growth perhaps arising from the gradual development of particle correlations as described by Toomre and Kalnajs (1991). Even in these cases, however, the amplitude rises more rapidly once $\delta \gtrsim 2\%$.

Thus even in these highly restricted simulations, spiral activity always exhibits runaway growth – albeit more and more delayed as N is increased – behavior that is quite clearly not in accord with linear theory predictions. The rapid growth once $\delta \gtrsim 2\%$ suggests that the behavior has already become nonlinear in some respect at this modest amplitude. Note that the largest number of particles, $N = 5 \times 10^8$, is within a factor of 100 of the number of stars in a real galaxy disk, where in reality the mass distribution is far less smooth, owing to the existence of star clusters and giant gas clouds.

¹⁰Interestingly, this idea harks back to the old “garden sprinkler” model for spirals proposed by Jeans (1923) (see also Jeans 1929, Fig. 55 & pp 357–360).



■ Fig. 18-10

The time evolution of the peak overdensity in a series of simulations of the half-mass Mestel disk with different numbers of particles. The model was predicted by Toomre (1981) to be globally stable. The ordinate shows the maximum values of δ on grid rings. The number of particles in each simulation rises by a factor 10 from $N = 5 \times 10^4$ to $N = 5 \times 10^8$, and the initial amplitude is $\propto N^{-1/2}$. Linear theory predicts the later amplitudes should have the same scaling

7.6 A Recurrent Instability Cycle?

Sellwood and Lin (1989) reported evidence for a recurrent instability cycle in their simulations of a very low-mass disk. They observed that each spiral pattern created substantial changes to the distribution of particles at the Lindblad resonances, which they suspected created conditions for a new global instability of the groove mode kind (see Sect. 6). They explicitly demonstrated that the later features were indeed true instabilities, since when they continued a parallel simulation after randomly shuffling the particles in azimuth at some moment, the pattern speed of the next mode to appear was the same as that in the original simulation that had not been interrupted.¹¹ Thus, each coherent wave leaves behind an altered DF that apparently provokes a new instability.

The runaway growth in the larger N models shown in Fig. 18-10 is not caused by a single unstable mode, but appears to be a succession of separate coherent waves, each having a lower rotation rate and reaching a higher amplitude than the last. Sellwood (2000) reported similar behavior in lower- N simulations, and demonstrated that one of the waves did indeed cause strong scattering at the ILR. It should be noted that resonance scattering is a second-order effect, but the evidence shown in Fig. 18-10 suggests that it becomes important at a relative overdensity of just $\Delta\Sigma/\Sigma \sim 2\%$. Exactly how the demise of one mode creates the conditions for the next instability remains unclear, however.

¹¹This behavior is inconsistent with the non-linear mode coupling ideas discussed in Sect. 7.4.

Since the only evidence for this behavior came from a (well-tested) N -body code, it seemed best not to pursue the idea further until supporting evidence could be found. Sellwood (1994) therefore expressed the hope that evidence of resonance scattering could be found in the HIPPARCOS measurements of the local stellar kinematics. The publication of the GCS with distances and full phase-space motions of $\sim 14,000$ F and G dwarf stars (see [Fig. 18-1](#)) enabled Sellwood (2010) to show that the so-called Hyades stream is in fact caused by scattering at a recent ILR.

This very recent evidence supports the idea that spirals in the local Milky Way, and presumably elsewhere, do in fact behave as the simulations had indicated. Further work is required to expose the details of the recurrence mechanism. However, it now seems misguided to search for a spiral instability as some devious sort of cavity mode in a smooth disk; i.e., the assumption of a featureless DF may have thrown the spiral baby out with the bathwater!

8 Buckling Instabilities and Warps

The optically visible parts of galactic disks are usually remarkably thin and flat, whereas the more extended H I disks of many edge-on galaxies appear noticeably warped with an integral sign shape (Sancisi 1976). Stellar warps (Reshetnikov et al. 2002; Saha et al. 2009) are much less pronounced than the warps in the extended gaseous disks. The long-known warp in the H I layer of the Milky Way (Oort et al. 1958) has been most recently analyzed by Levine et al. (2006), while the dust and stars of the outer disk appear to be distorted in a similar, though less extensive, shape (Reyl   et al. 2009). Both gaseous and stellar warps are frequently asymmetric, as appears to be the case for the Milky Way. The fact that stellar warps usually follow the same warped surface as do the gaseous ones (see also Cox et al. 1996) is strong evidence that warps are principally a gravitational phenomenon.

Warps are extremely common. H I observations of edge-on galaxies (Bosma 1991; Garc  a-Ruiz et al. 2002b) find very high fractions of warps, and the true fraction must be even higher, since warps directed close to our line of sight may be missed. Warps can also be detected kinematically even when the system is not edge-on. Their ubiquity suggests that warps are either repeatedly regenerated or long-lived.

Briggs (1990) studied a sample of 12 warped galaxies with high-quality 21-cm data, and found that galactic warps obey three fairly simple rules:

1. The H I layer typically is coplanar inside radius R_{25} , the radius where the B-band surface brightness is $25 \text{ mag arcsec}^{-2} = 25 \mu_B$, and the warp develops between R_{25} and $R_{26.5}$ (the Holmberg radius).
2. The **line of nodes** (LoN) is roughly straight inside $R_{26.5}$.
3. The LoN takes the form of a looselywound *leading spiral* outside $R_{26.5}$.

Kahn and Woltjer (1959) first drew attention to the winding problem presented by warps (see also BT08 Sect. 6.6.1) and argued that while self-gravity would slow the differential precession, it could not be strong enough to create a long-lived warp. Lynden-Bell (1965), on the other hand, suggested that warps could result from a persisting misalignment between the spin axis and the disk normal, i.e., a long-lived mode.

This section describes the theory of bending waves in general before addressing warps more directly. Unstable bending modes will be denoted **buckling modes**, although other names that come from plasma physics, such as hose, firehose, or hosepipe instabilities, are commonly used.

8.1 Local Theory of Bending Waves

Toomre (1966) considered the bending stability of an infinite, thin slab of stars having a velocity dispersion σ_x in the x -direction and some characteristic thickness z_0 . The self-gravity of the slab causes it to bend coherently provided the vertical oscillations of stars are adiabatically invariant as they move over the bend, which requires the slab to be thin compared with the wavelength of the bend. Furthermore, if the bending amplitude is small, its effect on the horizontal motion is negligible.

Toomre derived the dispersion relation for small-amplitude, long-wave ($kz_0 \ll 1$) distortions of the form $h(x, t) = He^{i(kx - \omega t)}$:

$$\omega^2 = 2\pi G\Sigma|k| - \sigma_x^2 k^2, \quad (18.13)$$

where Σ is the vertically integrated surface density of the slab. The first term on the RHS is the restoring force from the perturbed gravity caused by the bend while the second is the inertia term due to the vertical acceleration needed for stars to follow the corrugations. The inertia term is destabilizing and outweighs the gravitational restoring force when $\lambda < \lambda_J = \sigma_x^2 / G\Sigma$, causing the distortion to grow exponentially. The unstable range of the buckling instability is precisely complementary to that of the Jeans instability in the plane in the absence of rotation, which is unstable for wavelengths $\lambda > \lambda_J$ (Toomre 1966).

The dispersion relation (18.13) predicts a buckling instability, at sufficiently short wavelengths, for any razor-thin system. However, it does not hold for wavelengths shorter than or comparable to the actual vertical thickness of the slab. Since $z_0 \sim \sigma_z^2 / (G\Sigma)$, where σ_z is the vertical velocity dispersion, one expects that bending at all wavelengths will be suppressed when σ_z / σ_x exceeds some critical value, which Toomre (1966) estimated to be 0.30.¹²

Araki (1985) carried through a linear normal mode analysis of the infinite, isothermal slab (Spitzer 1942; Camm 1950), which has the properties $\rho = \rho_0 \operatorname{sech}^2(z/2z_0)$, $z_0 = \sigma_z^2 / (2\pi G\Sigma)$, and $\Sigma = 4\rho_0 z_0$. He assumed a Gaussian distribution of x -velocities, with $\sigma_x \neq \sigma_z$, and determined the range of unstable wavelengths as the slab was made thicker. He showed that the buckling instability could be avoided at all wavelengths when $\sigma_z > 0.293\sigma_x$, in good agreement with Toomre's earlier estimate. At the marginal stability threshold, the last unstable mode has a wavelength $\simeq 1.2\lambda_J$.

Galaxy disks are not, of course, infinite slabs subject to plane-wave distortions, but the radial velocity dispersion, σ_R , which is larger than the azimuthal dispersion, could perhaps be great enough to drive a buckling instability. As the observed ratio of velocity dispersions for stars in the Solar neighborhood is $\sigma_z / \sigma_R \sim 0.6$ (Nordström et al. 2004), Toomre (1966) concluded that at least this region of our Galaxy is "apparently well clear of this stability boundary." It should be noted that the approximate solution for the potential, which requires $kR \gg 1$, is stretched in this case, since $\lambda_J \approx 7 \text{ kpc}$ when $\sigma_R = 40 \text{ km s}^{-1}$ and $\Sigma = 50 M_\odot \text{ pc}^{-2}$.

When the disk is embedded in some external potential, arising from a halo or the distant bulge of the galaxy say, the dispersion relation for short-wavelength waves in a thin slab (18.13) acquires an additional stabilizing term

$$\omega^2 = v_{\text{ext}}^2 + 2\pi G\Sigma|k| - \sigma_R^2 k^2, \quad (18.14)$$

¹²Kulsrud et al. (1971) and Fridman and Polyachenko (1984) considered the bending instability in a constant density slab of stars with sharp edges.

where $v_{\text{ext}}^2 = |\partial^2 \Phi_{\text{ext}} / \partial z^2|_{z=0}$ (BT08 Eq. 6.114). Taking this additional factor into account further reinforces local stability and global, axisymmetric simulations (Sellwood 1996a) confirmed that Toomre's conclusion holds everywhere in an axisymmetric, but otherwise plausible, model of the Milky Way disk.

As for WKB spiral waves (☛ Sect. 3.2), the dispersion relation (☛ 18.14) can be generalized to *tightly wrapped* non-axisymmetric bending waves simply by replacing ω with $\omega - m\Omega$, with the angular rate of precession of the bending wave that has m -fold rotational symmetry being $\Omega_p = \omega/m$. It should be borne in mind that since observed warps in galaxies are very far from being tightly wound, analyses that make this approximation yield results that are at best only indicative of the dynamical behavior.

Vertical resonances between the bending wave and the stars arise where $m(\Omega_p - \Omega) = \pm v$ (☛ Sect. 2.6), although the meaning of v here depends on the context. BT08 (Sect. 6.6.1) considered only razor-thin disks, for which the internal oscillation frequency v_{int} is infinite and the resonances occur where $m(\Omega_p - \Omega) = \pm v_{\text{ext}}$. In disks of finite thickness, the stars have a natural internal vertical frequency $v_{\text{int}}^2 \approx 4\pi G\rho_0$ (exact in the mid-plane of an infinite slab), and for vertical resonances the appropriate value of $v = (v_{\text{int}}^2 + v_{\text{ext}}^2)^{1/2}$. Henceforth, v will be used to mean either this total frequency in a thickened layer or v_{ext} for a razor-thin sheet.

☛ Equation 18.14 is satisfied for both $\pm m(\Omega_p - \Omega)$, implying two possible pattern speeds known as **fast** and **slow waves**. Because the gravity term raises the Doppler-shifted frequency above v , waves in a cold disk ($\sigma_R = 0$) extend away from the vertical resonances, and do not exist in the region between them that includes CR. The slow wave, which is a retrograde pattern for $m = 1$, is of most interest because the unforced precession rate, $\Omega_p = \Omega - v$, has much the milder radial variation.

Much like density waves, neutrally stable bending waves propagate in a disk in the radial direction at the **group velocity** (Toomre 1983)

$$v_g = \frac{\partial \omega}{\partial k} = \frac{\text{sgn}(k)\pi G\Sigma - k\sigma_R^2}{m(\Omega_p - \Omega)}. \quad (18.15)$$

Since the denominator is negative for slow bending waves, trailing waves ($k > 0$) in a cold disk ($\sigma_R = 0$) propagate inward, while leading waves ($k < 0$) propagate outward. As for waves in nonuniform whips, the bending amplitude rises when the wave propagates into a region of lower surface density, and vice versa.

Unfortunately, a full description of wave propagation in a sheet of finite thickness requires a solution of the linearized Boltzmann and Poisson equations (Toomre 1966; Araki 1985; Weinberg 1991; Toomre 1995). The results are not analytic and surprisingly more complex than (☛ 18.14) for the razor-thin case. Because the vertical potential of the disk is anharmonic, stars whose vertical oscillation takes them far from the mid-plane have lower vertical frequencies. Thus, any Doppler-shifted frequency $m(\Omega_p - \Omega)$ of the bending wave will be in resonance with some stars that will damp the wave by converting wave energy into increased random motion. In general, short-wavelength modes $kz_0 \gtrsim 0.5$ damp in less than one wavelength (Weinberg 1991; Toomre 1995), while long-wavelength modes $kz_0 \ll 1$ can propagate large distances.

Sellwood et al. (1998) studied the propagation and damping of axisymmetric bending waves in disks having both finite thickness and nonzero random velocities. Waves launched from the center of the disk at a fixed frequency propagated outward and were damped as they approached the vertical resonance. As a result the disk thickened over a small radial range, with the peak occurring just interior to the resonance.

8.2 Global Bending Modes

Because of the complication caused by finite thickness, analytic work almost always adopts the razor-thin approximation, which also requires $\sigma_R = 0$, since thin disks with velocity dispersion are buckling unstable.¹³ With these assumptions, the disk can be approximated by a finite number of gravitationally coupled circular rings of matter, each having the appropriate angular momentum, which always admits a discrete spectrum of bending modes. The real modes of the continuum disk can be distinguished by showing they are independent of the number of rings employed.

The gravitational restoring forces are correctly captured in this approach, but the lack of random motion omits the additional coupling between adjacent mass elements caused by the epicyclic motions of the stars. This extra mechanism further stiffens the disk's resistance to bending (Debattista and Sellwood 1999), especially in high-density inner regions where radial epicycles are larger.

Hunter and Toomre (1969) developed the coupled ring approach to study the bending dynamics of rotationally supported, razor-thin disks with no random motion and no halo. They were able to prove that general disks of this kind have no axisymmetric ($m = 0$) or warping ($m = 1$) instabilities, but they could not extend their proofs to higher sectoral harmonics. In fact they noted that were the disk composed of two equal counter-rotating populations of stars (on circular orbits) it would be buckling unstable for all $m \geq 2$.

Hunter and Toomre (1969) also studied the particular case of the sharp-edged Maclaurin disk in which all stars orbit at the same angular rate. When the sense of rotation is the same for all stars, the disk is stable to all bending waves, and there is a simple set of discrete neutral modes for all m . Polyachenko (1977) extended his analysis to Maclaurin disks with random motion (the Kalnajs disks) and was able to solve for the complete spectrum of bending modes. Needless to say, the introduction of random velocities in this razor-thin system caused buckling instabilities to appear for all m .

Since the Maclaurin disk bears little resemblance to real galaxies, Hunter and Toomre (1969) modified the disk to blur the sharp edge. They demonstrated that discrete warp modes in a cold, razor-thin disk can exist only when the edge is unrealistically sharp. Note that all isolated disks admit two trivial zero-frequency modes: a vertical **displacement** of the entire disk and a **tilt** of the disk plane to its original direction. More interesting standing wave solutions (modes) require traveling waves to reflect off the disk edge, but a realistic disk with a fuzzy edge does not reflect bending waves, because the group velocity decreases with the disk surface density (► 18.15) and a wave packet in a cold disk will never reach the edge (Toomre 1983).

8.3 Simulations of Buckling Modes

Sellwood and Merritt (1994) used N -body simulations to study the global instabilities of hot disks with no net rotation, i.e., with half the particles counter-rotating. (See ► Sect. 9.7 for buckling modes of more normal disks with large net rotation.) The form and vigor of the principal instabilities in any one of their models varied with the balance between radial and azimuthal pressure and with disk thickness: an in-plane lop-sided instability was the most disruptive for cool models (► Sect. 5). The radially hotter thin disks were disrupted by axisymmetric bending

¹³A razor-thin disk is not destabilized by orbital motions with no velocity spread.

instabilities (bell modes), as illustrated in [Fig. 18-11](#). The instability creates a thick inner disk resembling a pseudo-bulge (Kormendy and Kennicutt 2004). Very thin disks also buckled in an $m = 2$ “saddle” mode, and an $m = 1$ warp instability was detectable in some models, but never dominant.

Remarkably, instabilities in a counter-streaming model having intermediate radial pressure caused rather mild changes and led to an apparently stable, moderately thin, and almost axisymmetric disk. The in-plane velocities in this model resemble those reported by (Rubin et al. 1992) for the S0 galaxy NGC 4550, and indicate this galaxy could be stable even without large quantities of dark matter. The stability of this end product also demonstrated that thin axisymmetric systems with modest radial pressure are more stable to bending modes than those having isotropic or radially biased DFs.

Merritt and Sellwood (1994) proposed a criterion for the stability of a stellar system to buckling modes that can be applied globally. A particle moving at speed \dot{x} in a mildly bent sheet with characteristic wavenumber k experiences a periodic vertical forcing at frequency $k\dot{x}$ and, like any harmonic oscillator, the phase of its response depends upon whether the forcing frequency is greater or less than its natural vertical frequency. If $k\dot{x} < \nu$, the density response of the system to an imposed perturbation will be supportive, and the disturbance can be sustained or even grow. However, if $k\dot{x} > \nu$ for most particles, the overall density response to the perturbation will produce a potential opposite to that imposed and the disturbance will be damped. Thus short waves are stable. They showed that their proposal successfully accounted for their N -body results, for the behavior of the infinite slab at short wavelengths, and for the apparent absence of elliptical galaxies more flattened than E7.

Sellwood (1996b) found a long-lived bending oscillation in an N -body simulation of a warm disk that was constrained to be axisymmetric. Apparently the system was able to support a standing wave, at most mildly damped, between the center and the edge of the disk, even though the stars had random motion in a disk of finite thickness. The frequency of the bending or flapping mode was low enough to avoid vertical resonances with almost all disk particles,

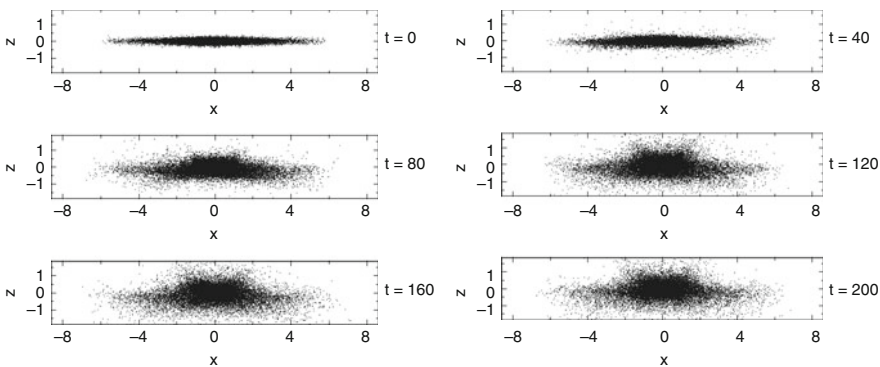


Fig. 18-11

The development of a buckling instability in a simple model of an isolated stellar disk with $Q \sim 2$. The disk is a KT/4 model described by Sellwood and Merritt (1994), with equal numbers of particles orbiting in each direction, but here the simulation uses a code of much higher spatial resolution. The buckling mode is axisymmetric and while non-axisymmetric features were permitted by the code, none developed. The disk mass and radial scale are unity, and $G = 1$; the orbit period at $R = 2$ is 16 in these units

in agreement with the requirement stated by Mathur (1990). While this result remains an isolated curiosity, since the flapping mode would probably be quickly damped in a halo, it is a clear counterexample to the argument by Hunter and Toomre (1969) that realistic disks do not possess global bending modes, which apparently holds only for disks without random motion.

8.4 Disks in Halos

Since the paper by Hunter and Toomre (1969), ideas of warp formation have relied in some way on the interaction between the disk and its dark matter halo. Dekel and Shlosman (1983) and Toomre (1983) suggested that a flattened halo misaligned with the disk could form a long-lasting warp. Sparke and Casertano (1988) and Kuijken (1991) obtained long-lived warps (dubbed **modified tilt modes**) of disks in rigid, misaligned halos, which were insensitive to the details of the disk edge. Lovelace (1998) studied the tilting dynamics of a set of rings also in a rigid halo, but assumed that the inner disk lay in the symmetry plane of the spheroidal halo.

Dubinski and Chakrabarty (2009) noted that the dark matter halos that result from cosmic structure formation simulations are usually aspherical, with frequent misalignments between the principal axes of the inner and outer halo. The disks in their simulations warped nicely when forced with slowly rotating, but otherwise rigid, perturbing fields representative of such halos.

However, dark matter halos are not rigid, and a responsive halo alters the dynamics in several ways. Nelson and Tremaine (1995) showed that were the inner disk misaligned with the principal plane of the flattened halo, as supposed by Sparke and Casertano (1988), its precession would be damped through dynamical friction, bringing the disk into alignment with the halo on time scales much shorter than a Hubble time. But a more compelling objection to the modified tilt mode emerged from N -body simulations with live halos: Dubinski and Kuijken (1995) found that the warp did not survive while Binney et al. (1998) showed that the inner halo quickly aligns itself with the disk, not vice versa. The large store of angular momentum in the disk maintains its spin axis, but the pressure-supported inner halo can readily adjust its shape slightly to align itself with the disk.

8.5 Misaligned Infall

The idea that galaxy warps are manifestations of *eternal* warp modes seems doomed by the damping effect of a live halo. But slowly evolving warps remain viable, provided that suitable external perturbations occur in enough cases.

In hierarchical galaxy formation scenarios, late infalling material probably has an angular momentum axis misaligned with the disk spin axis. Ostriker and Binney (1989) therefore proposed that warps arise due to the slewing of the galactic potential as material with misaligned angular momentum is accreted. Structure formation simulations by Quinn and Binney (1992) confirmed that the mean spin axis of a galaxy must slew as late arriving material rains down on the early disk. The less-than-critical matter density in modern Λ CDM universe models implies that infall is less pervasive at later times, but it manifestly continues to the present day in gravitationally bound environments (Sancisi et al. 2008, and chapter by van Woerden and Bakker).

Jiang and Binney (1999) and Shen and Sellwood (2006) presented results of experiments in which a disk was subjected to the torque from a misaligned, massive torus at a large radius.

This well-defined perturbation is a very crude model of an outer halo that is rotationally flattened, and having its spin axis misaligned with that of the disk. It is misaligned and farther out because, in hierarchical scenarios, the mean angular momentum of the later arriving outer halo is probably greater and misaligned from that of the original inner halo and disk. The accretion axis is, in reality, likely to slew continuously over time, so a model with a constant inclination is somewhat unrealistic.

Rather than striving for realism, Shen and Sellwood (2006) used this simple forcing to reach an understanding of how the warp develops and why the LoN usually forms a loosely wound leading spiral. The inner disk maintained a coherent plane because of both self-gravity and random motion. The torque arising from the misaligned outer torus caused the inner disk, and the aligned inner halo, to precess rigidly even though the torque increased with radius, but the outer disk beyond $\sim 4R_d$ started to warp.

As soon as the outer disk became misaligned with the inner disk, the strongest torque on the outer parts of the disk arose from the inner disk. The torque from the interior mass was responsible for the leading spiral of the line of nodes, even though the adopted external field would have produced a trailing spiral. The fact that the LoN of most warps forms a leading spiral over an extended radial range seems to imply massive disks.

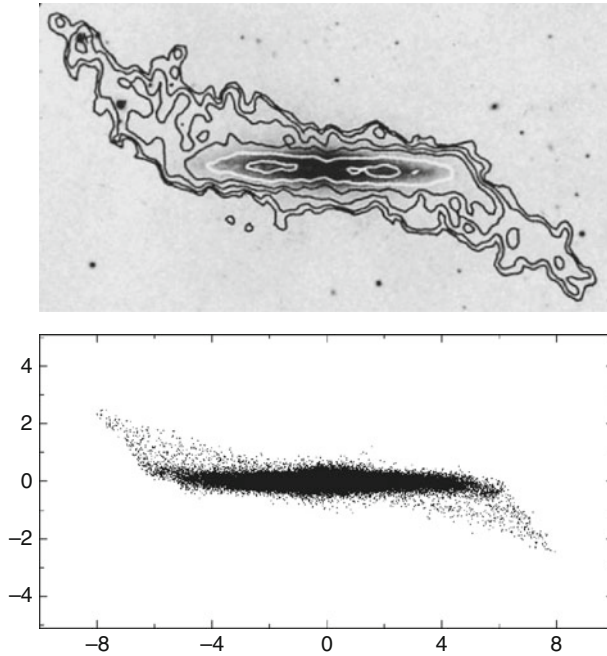
Even though the disk precessed due to the external torque, its motion was barely damped over many Gyr, in contrast to the expectations from Nelson and Tremaine (1995). Damping was weak because the slow precession rate allowed the inner halo to remain closely aligned with the disk, which therefore caused little drag. The weak damping seemed to be caused more by the relative precession of the inner and outer parts of the halo. Also the warp evolved slowly as the layer settled to the main plane at gradually increasing radii, in apparent agreement with the decreasing outward group velocity.

🔗 *Figure 18-12* shows the H 1 observation of NGC 4013 by Bottema (1996), together with the warp obtained in the simulation by Shen and Sellwood (2006). Their simulations revealed that the warp persisted for cosmologically interesting times, even when the external forcing field was removed. Thus the persistence of warps is not nearly as perplexing as previous studies had suggested. Furthermore, the model had a flat inner disk and the warp in the outer disk matched all of Briggs's rules quite well.

A fixed outer torus is clearly unrealistic and the halo axis probably shifts continuously or episodically, as argued by Quinn and Binney (1992), making warp lifetimes a side issue. Warps formed this way can be repeatedly regenerated when a new infall event happens. Since cosmic infall and mergers are more likely to happen in a denser environment, warps can be induced more frequently in such an environment, which is consistent with the statistics (García-Ruiz et al. 2002b).

8.6 Warps Driven by Tides

Tidal interactions between galaxies could be an additional cause of warps in disks, and one that is quite likely to produce asymmetrical warps. This idea has been explored most fully to explain the warp of the Milky Way's (MW) disk that results from the proximity of the Magellanic Clouds, especially the Large Magellanic Cloud (LMC). The orientation of the principal axes of the warp (Levine et al. 2006) at least seems favorable to this hypothesis. Bailin (2003) suggested that the Sagittarius dwarf galaxy is another possible culprit.



■ Fig. 18-12

The upper panel shows the observed H I warp of NGC 4013 (Bottema 1996), and the lower panel the warp in the simulation by Shen and Sellwood (2006) that closely resembles it. The length unit on the axes is the scale length R_d of the exponential disk

Hunter and Toomre (1969), in models that did not include dark matter, concluded that the LMC could be responsible for the warp of the MW, but only if it was a lot more massive than was then suspected and had recently passed close to the edge of the disk. Garcia-Ruiz et al. (2002a) tested the hypothesis in fully self-consistent simulations that included live halos and used updated information about the distance and motion of the LMC. They concluded that neither the amplitude, nor the orientation, of the warp in the disk of the MW was consistent with the tidal hypothesis. Weinberg and Blitz (2006), on the other hand, suggested that a cooperative response from the halo to the passage of the LMC could have generated the observed warp in the disk. Thus there is no consensus yet on the origin of the warp in the Milky Way.

9 Bars

Bars are common features in disk galaxies. An earlier review (Sellwood and Wilkinson 1993) of the vast topic of bars is now somewhat dated but, as it provides a still useful summary of the basic dynamics, the present article will update a few main points and the reader is referred to the earlier review for a more detailed discussion.

9.1 Origin of Bars

For a long time, dynamicists were struggling to understand the absence of bars in some disk galaxies (Ostriker and Peebles 1973), but since reasonable models of luminous galaxies that include a dense bulge of moderate mass are now known to be stable (▶ Sect. 4), the problem has become almost the opposite! Strong bars are seen in many galaxies whose mass distributions now appear unfavorable to the dynamical bar instability, as evidenced by a nuclear gas ring (see ▶ Sect. 9.4), which can form only if the center is dense enough to have inhibited bar formation by Toomre's mechanism.

However, the fact that the center is dense today does not require that it was dense when the bar formed; secular inflow of gas (▶ Sect. 9.4) can build up the central density after the bar has formed. Alternatively, the bar could have grown in size and slowed through disk evolution (Sellwood 1981; Berentzen et al. 2007) or halo friction (Athanasoula 2002, see Sect. 9.8), or have been triggered by large density fluctuations in the disk (Sellwood 1989a), by tidal interactions (Noguchi 1987; Berentzen et al. 2004; Curir et al. 2006), halo substructure (Romano-Díaz et al. 2008b) or a non-axisymmetric halo (Dubinski and Chakrabarty 2009). Any of these considerations, or that described in the next paragraph, could plausibly reconcile the existence of bars in these galaxies with Toomre's stabilizing mechanism.

In an elegant piece of dynamics, Lynden-Bell (1979) demonstrated that the inner parts of galaxies are regions where eccentric orbits have a tendency to align themselves, which allows a bar to grow slowly through orbit trapping. The region where a cooperative response to a mild perturbation occurs is where the overall radial density profile of the galaxy flattens into a more uniform core. In this region, the radial variation of $\Omega - \kappa/2$ has a maximum at some nonzero radius and an infinitesimal bar pattern can have a pattern speed that allows two ILRs. Lynden-Bell's aligning mechanism, which operates best on the more eccentric orbits, requires $\Omega_p \simeq \Omega_\phi - \Omega_R/2$ of orbits in the aligning region. As this pattern speed is much lower than that expected from the global bar instability (▶ Sect. 4.1), the aligning mechanism offers an additional route to bar formation in otherwise globally stable disks. Although the cooperative region has a small radial extent, Lynden-Bell (1979) suspected the bar could be much larger.

Erwin (2005) presented a useful study of bar properties and pointed out that bars in late-type galaxies are often much smaller relative to the disk size than are those formed in simulations.

9.2 Frequency of Bars

Strong bars are clearly visible in 25–30% of disk galaxies (e.g., Masters et al. 2010), and the fraction rises to $\gtrsim 50\%$ when more objective criteria are applied to red or near-IR images (Eskridge et al. 2000; Marinova and Jogee 2007; Reese et al. 2007, and further references cited below). Barazza et al. (2008) found a higher bar fraction in later Hubble types, while Méndez-Abreu et al. (2010) found no bars in either very luminous or very faint galaxies.

Whatever may be the mechanism responsible for the formation of bars in real galaxies, none of the above suggestions makes a clear prediction for the frequency of bars. Bosma (1996), Courteau et al. (2003), and others have pointed out that barred galaxies seem little different from their unbarred cousins in most respects, e.g., they lie on the same Tully-Fisher relation. Minor systematic differences do exist: For example, Davoust and Contini (2004) note that barred galaxies seem to have smaller mass fractions of neutral H 1 gas, but this seems more likely to be the result, rather than the cause, of the bar. The anti-correlation of bar frequency

with the bulge half-light (Barazza et al. 2008) possibly results from Toomre's stabilizing mechanism, but this cannot be the whole story because some near-bulgeless disks are unbarred while other barred disks have massive bulges.

If no dynamical property, other than their eponymous one, can be identified that cleanly separates barred from unbarred galaxies, then the existence of a bar in a galaxy may possibly be determined by external factors, such as a chance encounter. Elmegreen et al. (1990) reported an increased fraction of bars in groups and clusters, but more recent work (Barazza et al. 2009; Li et al. 2009; Aguerri et al. 2009) has found little or no variation of bar fraction with environment. It is also possible that the bar fraction could be changing with time; different groups disagree (Jogee et al. 2004; Sheth et al. 2008), probably because observations of galaxies at significant look-back times are subject to systematic difficulties due to band-shifting and changing spatial resolution (see also Elmegreen et al. 2007).

A radical alternative is to regard bars as transient features that form and decay, and that the current fraction of barred galaxies represents the duty cycle (Bournaud et al. 2005). Bars in N -body simulations are dynamically rugged objects that appear to last indefinitely. Of course, they could be destroyed by a merger event, e.g., although not much in the way of a cool disk would survive such an event. Norman and coworkers (Pfenniger and Norman 1990; Norman et al. 1996) have argued that bars can be destroyed by the accumulation of mass at their centers, which may lead to a pseudo-bulge and/or a hot inner disk. However, Shen and Sellwood (2004) and Athanassoula et al. (2005) found that unreasonably large and/or dense mass concentrations were required to cause their bars to dissolve. The simulations by Bournaud et al. (2005) uniquely show that gas accretion may aid the dissolution of the bar and, with star formation, could recreate a cool disk that would be needed to make a new bar. Even if this behavior can be confirmed by others, the model requires very substantial gas infall. Moreover, the continued existence of the hot old disk and the buildup of a dense center makes every cycle of this speculative picture harder to achieve.

In the distant future, galaxy formation simulations may have the quality and resolution perhaps to be able to predict the correct bar fraction, and thereby reveal their cause.

9.3 Structure of Bars

This section gives a brief description of a few important aspects of the orbital behavior in large-amplitude bars, and the reader is referred to Sellwood and Wilkinson (1993) for a more comprehensive discussion. Weak bars can be treated using epicycle theory (BT08 Sect. 3.3.3). Most early orbit studies in strongly barred potentials considered motion confined to the plane perpendicular to the rotation axis. Even though 3D motion is much richer, the fundamental structure of bars is most easily understood from in-plane orbits.

Since bars are believed to be steadily rotating, long-lived objects, it makes sense to discuss their structure in a frame that corotates with the potential well at the angular rate Ω_p . A rotating frame has the **effective potential**

$$\Phi_{\text{eff}} = \Phi - \frac{1}{2}\Omega_p^2 R^2, \quad (18.16)$$

where Φ is the potential in an inertial frame.

The effective potential surface in the disk plane (● 18.16) resembles a volcano, with a central crater, a rim, and a steeply declining flank. The crater is elongated in the direction of the bar,

and the rim has four **Lagrange points**: two maxima, L_4 and L_5 , on the bar minor axis and two saddle points, L_1 and L_2 , on the bar major axis. (The fifth Lagrange point, L_3 , is the local potential minimum at the bar center.) Because of Poisson's equation, the density contours of the bar must be more elongated than those of the inner Φ_{eff} .

Neither E nor L_z is conserved in non-axisymmetric potentials, but Jacobi's invariant I_J (◆ 18.6) is conserved even for strong bars that rotate steadily. Since $I_J = \frac{1}{2}|\mathbf{v}|^2 + \Phi_{\text{eff}}$, where \mathbf{v} is the velocity in the rotating frame, contours of Φ_{eff} bound the possible trajectories of stars having I_J less than that contour value. Stars that are confined to the bar, which also have $I_J < \Phi_{\text{eff}}(L_1)$, are of most interest here.

A **periodic orbit** is a possible path of a star in the rotating frame that retraces itself, usually after a single turn around the center, but always after a finite number of turns. Because the orbits close, the orbital period in the rotating frame is commensurable with the radial period and these orbits are also described as resonant orbits of the (strongly non-axisymmetric) potential.

Periodic orbits can be either **stable**, in which case a star nearby in phase space oscillates (librates) about its **parent** periodic orbit in an epicyclic fashion, or they are unstable, in which case the trajectory of a star nearby in phase space diverges exponentially (at first) from the periodic orbit. The orbits of stars that librate around a periodic orbit are known as **regular** orbits, and those that do not librate about any periodic orbit are known as **irregular** or **chaotic** orbits. Chaotic orbits have only a single integral, I_J , while regular orbits have an additional integral (two more in 3D) that confines their motion to a hypersurface of smaller dimension in phase space. Regular orbits are the more interesting because the star's orbit can be more elongated than the potential surface that confines it, which is of great value when building a self-consistent bar model.

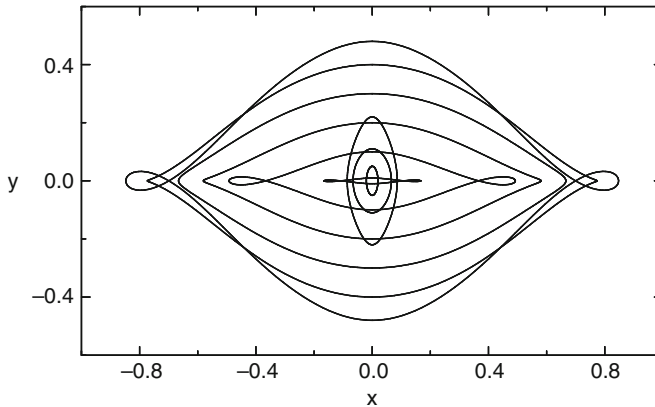
The main features can be illustrated in the simple potential (cf. BT08 Eq. 3.103)

$$\Phi_{\text{eff}}(x, y) = \frac{1}{2}v_0^2 \ln \left(1 + \frac{x^2 + y^2/q^2}{R_c^2} \right) - \frac{1}{2}\Omega_p^2 R^2, \quad (18.17)$$

where $R^2 = x^2 + y^2$, R_c is a core radius inside of which the potential is approximately harmonic, $q \leq 1$ is the flattening, and v_0 is the circular speed at large R when $q = 1$. As in BT08, the values are: $v_0 = 1$, $q = 0.8$, $R_c = 0.03$, and $\Omega_p = 1$. With these parameters, the major-axis Lagrange points lie at a distance $R_L \simeq 0.9996$ from the bar center.

The 2:1 resonant periodic orbits shown in ◆ Fig. 18-13 have a range of I_J values, but all close in the rotating frame after two radial oscillations for every turn about the center. Orbits of the main family, denoted x_1 , are elongated parallel to the bar (horizontal in the Figure), and are referred to as the **backbone of the bar**, since the majority are stable. Stars on these orbits all move in the same direction as the bar rotates, but have shorter periods because they are interior to the Lagrange points (or, more loosely, they are inside CR). Some x_1 orbits are simple closed figures, but others have loops near the outer ends where the star progresses around the center of the galaxy more slowly than does the rotating frame. Sparke and Sellwood (1987) and Voglis et al. (2007) found that a large fraction of particles in an N -body bar librate around them, and the same behavior is expected for stars in real bars.

Another important family of 2:1 orbits appears near the center of the bar, but is elongated perpendicular to the bar axis, as also shown in ◆ Fig. 18-13. This family, denoted x_2 , is almost always stable and is present in many realistic bar potentials. Since Lindblad resonances (◆ Sect. 2.4) are defined for infinitesimal perturbations to axisymmetric potentials, the label ILR is a very loose usage in strong bars. It is true that the x_2 family appears in barred potentials only,



■ Fig. 18-13

Examples, in a rotating bar potential, of important periodic orbits that close after two radial oscillations for every turn about the center, the 2:1 resonant families. Those orbits elongated parallel to the bar axis (horizontal) are members of the x_1 family. The x_2 orbits are elongated perpendicular to the bar

but not always, when the axisymmetric mass distribution and pattern speed admit one or more ILRs. Furthermore, x_2 orbits orient themselves perpendicular, while the x_1 family is parallel, to the bar axis, which is directly analogous to the abrupt phase shift that occurs in the response of a forced harmonic oscillator as the forcing frequency crosses its natural frequency (*cf.* Sanders and Huntley 1976). At finite amplitude, the near circular orbit sequence acquires gaps at the 2:1 resonance bifurcations, which become broader as the bar amplitude rises. Thus the radial extent of the x_2 family shrinks as the strength of the bar is increased, and it may disappear entirely. Even if one is careful to say that the appearance of the x_2 family is the generalization of the ILR to strong bars, the radius of this resonance is still badly defined because the orbits that appear inside it can be quite eccentric.

As I_J approaches the value of $\Phi_{\text{eff}}(L_1)$, the time to complete a full turn in the rotating frame lengthens and additional orbit families appear. Orbits that close after any number of radial oscillations can be found in principle, but of these only the 4:1 resonant orbits are of dynamical significance to bar structure. As the period lengthens, the proliferation of orbit families causes a precipitous decrease in the stable regions around each parent and chaotic behavior ensues (Chirikov 1979). The onset of chaos near CR led Contopoulos (1980) to expect the density of a self-consistent bar to drop steeply near the major axis Lagrange points, leading to the rule that the length of a bar is limited by CR. This rule predicts that the parameter (Elmegreen 1996)

$$\mathcal{R} \equiv R_L/a_B > 1, \quad (18.18)$$

where a_B is the semi-major axis of the bar. In principle, self-consistent bar dynamics could allow bars with $\mathcal{R} \gg 1$, although empirical bar pattern speed estimates (► Sect. 9.5) mostly find that CR is in fact only slightly beyond the end of the bar.

There are many more in-plane orbit families, but few are of dynamical importance to the structure of the bar. See Sellwood and Wilkinson (1993) for a fuller account.

The extension to 3D allows for many more resonances between the in-plane motion and the vertical oscillations. While there is a rich variety of behavior (Pfenniger and Friedli 1991; Skokos et al. 2002), the backbone x_1 family from 2D continues to be the most important, but now with a “tree” of orbits also librating vertically. The new periodic orbits that appear in 3D have similar projected shapes as the in-plane x_1 family, but they also librate vertically a small number of times over the same period as the motion in the plane. Patsis et al. (2002) highlighted the orbit families that they found to be of importance for the “boxy” appearance of edge-on bars (see also ▶ Sect. 9.7).

9.4 Gas Flow

When pressure and magnetic forces can be neglected, any mild dissipation will drive gas to move on stable periodic orbits. An organized streaming gas flow pattern is expected wherever the simplest periodic orbits over a range of energies can be nested and intersect neither with neighboring orbits, nor with themselves. Shocks, where pressure ceases to be negligible, must occur in flows either where periodic orbits self-intersect or where gas flows on two separate orbits cross. ▶ Figure 18-13 shows that were gas to flow in that adopted bar potential, shocks would be inevitable because many orbits self-intersect (the loops) and, in particular, x_2 orbits cross the x_1 family. Thus shocks are a general feature of cool (low pressure) gas flows in bars.¹⁴

Full hydrodynamic simulations (e.g., Roberts et al. 1979; Athanassoula 1992; Fux 1999) are needed to determine the flow pattern. Shocks are offset to the leading sides of the bar major axis in models having reasonable parameters. Prendergast (1962) seems to have been the first to associate the dust lanes in bars with the locations of shocks in the gas.

Shocks convert some kinetic energy of bulk motion in the gas into heat, which is radiated efficiently. Furthermore, the offset location of the shock causes the gas to spend more than half its orbit on the leading sides of the bar, where it is attracted backward toward the bar major axis, causing it to lose angular momentum.¹⁵ Thus gas in the bar region must settle a little deeper into the potential well on each passage through a shock, i.e., the bar drives gas inwards, the angular momentum it loses being given to the bar.

The inflow stalls where gas settles onto the x_2 orbit family, which is found in bar models that have dense centers, leading to a build up of gas. This behavior can be associated with nuclear rings of dense gas (Regan et al. 2002), which are often the sites of intense star formation also (Maoz et al. 2001; Benedict et al. 2002). If this were the whole story, the gas could not be driven any further inward, but there is both observational evidence, in the form of spiral dust lanes and star formation (e.g., Carollo et al. 1998), and some theoretical work (Wada 2001) to suggest that self-gravity causes inflow to continue. However, the existence of high gas density and rapid star formation in the nuclear ring indicates that only a small fraction of the gas continues inward.

Modeling the gas flow in a specific galaxy allows one to determine two properties of the bar that are hard to constrain otherwise. Estimating the gravitational potential of the galaxy from a photometric image plus a dark halo, Weiner et al. (2001) fitted for both the disk mass-to-light (M/L) and Ω_p in the galaxy NGC 4123. These authors also described the procedure in detail. Results for a number of other galaxies were reported by Pérez et al. (2004), Weiner (2004), and Zánmar Sánchez et al. (2008), although these last authors were unable to obtain

¹⁴Shocks may be avoided when pressure is significant (Englmaier and Gerhard 1997).

¹⁵The opposite happens in shocks outside CR, where the gas gains angular momentum from the bar.

an entirely satisfactory fit. Lindblad et al. (1996) fixed the M/L and fitted only for Ω_p . All these fits preferred rapidly rotating bars in heavy disks. Pérez (2008) confirmed that the best fit parameters of M/L and Ω_p were the same for both 2D Eulerian (Godunov) and 3D Lagrangian (SPH) hydrodynamic methods.

Even though the quadrupole field of a bar decays quickly with radius, it can be strong enough to drive a spiral shock in the gas of the outer disk, as originally demonstrated by Sanders and Huntley (1976). Schwarz (1981) showed that when gas is modeled as inelastic particles, it is driven outward to form an outer ring (see review by Buta and Combes 1996). However, it is unclear that spirals in the outer disks of real barred galaxies are the responses to the bar, and they may owe more to self-excited structures than to bar driving (Sect. 7, Sellwood and Sparke 1988; Buta et al. 2009). In addition, the outer spiral response to an imposed bar is not steady in modern simulations, with the shapes of the driven arms cycling through a broad range. For these reasons, the gas flow models fitted to individual galaxies should rely primarily on the fit within the bar and pay little attention to the outer disk.

Kranz et al. (2003) tried a similar approach, but fitted a spiral pattern instead of a bar, which may yield unreliable results for two reasons: (1) The lifetimes of spiral arms are believed to be short (Sect. 7), leading to broader resonances and stronger gas responses than would arise in simulations that assume a slowly evolving pattern, and (2) the observed spirals could be the superposition of several features with different angular rotation rates. As bars undoubtedly last longer than do spirals and dominate the non-axisymmetric potential, fits to bars in galaxies are likely to yield better disk mass estimates.

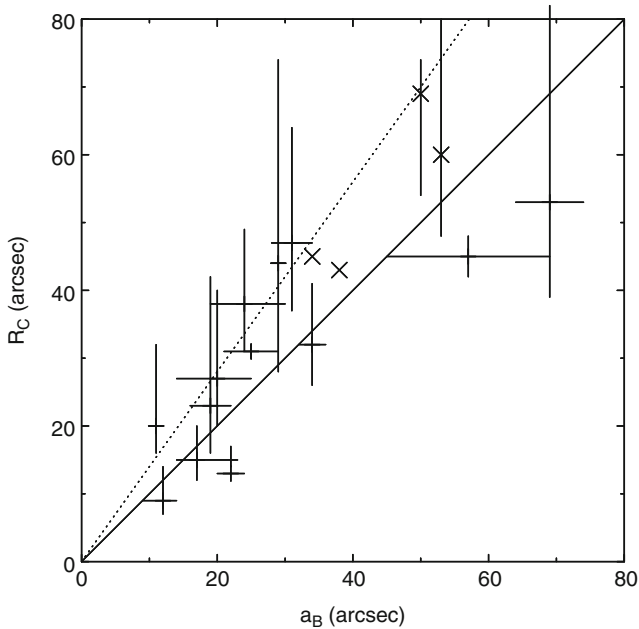
9.5 Bar Pattern Speeds

Tremaine and Weinberg (1984a) devised a method to measure the pattern speed of a bar directly from observations of a tracer component, which must obey the equation of continuity. Their original method assumes that the galaxy has but a single pattern, and would yield a misleading result were there more than one pattern, each rotating at a different angular rate.

The stellar component of early-type barred galaxies is believed to obey the equation of continuity because these galaxies have little dust obscuration and no star formation. They also rarely possess prominent spirals in the outer disk. Results of many studies using this method for early-type barred galaxies were summarized by Corsini (2008) and are shown in Fig. 18-14. While some individual measurements are quite uncertain, the data seem to favor $1 < \mathcal{R} \lesssim 1.4$. Chemin and Hernandez (2009) found a counterexample in a low-luminosity galaxy.

Fathi et al. (2009) and Meidt et al. (2009) applied the method of Tremaine and Weinberg (1984a) to ionized and molecular gas, respectively. Both groups argue that this is valid, even though the separate gas components do not obey the continuity equation that underlies the method. Fathi et al. (2009) generally find fast bars. Meidt et al. (2008) generalized the method to attempt to measure radial variations in the pattern speed and Meidt et al. (2009) found suggestions of pattern speeds that are lower at large radii than those near the center.

Other methods can yield indirect estimates of bar pattern speeds. Fits of models of the gas flow (Sect. 9.4) have been reported for a few galaxies, finding $\mathcal{R} \sim 1.2$. Athanassoula (1992) argued that the shapes and locations of dust lanes in bars also seem to suggest that $\mathcal{R} \simeq 1.2$. Associating a ring in a barred galaxy with the location of a major resonance with the bar (Buta and Combes 1996) yields, with kinematic information, an estimate of the pattern speed.



■ Fig. 18-14

Summary of direct pattern speed measurements for bars collected by Corsini (2008). The *diagonal line* shows $\mathcal{R} = 1$ and the *dotted line* $\mathcal{R} = 1.4$. The crosses mark values for which error bars are unavailable

Rautiainen et al. (2008) computed models of the stellar and gas (sticky particles) responses to forcing by photometric models of 38 barred galaxies, in which they assumed that the entire non-axisymmetric structure rotated at the same pattern speed. They attempted to match the model to the visual morphology of the galaxy, and found a range of values for \mathcal{R} . However, in most cases where $\mathcal{R} \gg 1$, the fit is dominated by the outer spiral, which may have a lower angular speed than does the bar.

9.6 Bars Within Bars

Erwin and Sparke (2002) and others have found inner **secondary bars** within the inner parts of >25% of large-scale or **primary bars**. They reported that the secondary bar has a length some ~12% of that of the primary bar and the deprojected angles between the principal axes of the two bars appeared to be randomly distributed, suggesting that the two bars may tumble at differing rates. This inference was supported by Corsini et al. (2003), who used the Tremaine and Weinberg (1984a) method to show that the two bars in NGC 2950 could not have the same rotation rates; Maciejewski (2006) used their data to argue that the secondary bar has a large retrograde pattern speed.

The theoretical challenge presented by these facts is substantial, and progress to understand the dynamics has been slow. Maciejewski and Sparke (2000) studied the orbital structure in a

potential containing two non-axisymmetric components rotating at differing rates. However, it is almost certain that the secondary bar can neither rotate at a uniform rate (Louis and Gerhard 1988) nor can it maintain the same shape at all relative phases to the primary.

Friedli and Martinet (1993) argued that gas was essential to forming secondary bars (see also e.g., Heller et al. 2001; Englmaier and Shlosman 2004). However, some of the collisionless simulations reported by Rautiainen and Salo (1999) and Rautiainen et al. (2002) manifested dynamically decoupled inner structures when the inner disk had high orbital frequencies due to a dense bulge. The structure was more spiral like in some models, but others appeared to show inner bars that rotated more rapidly than the main bar.

Debattista and Shen (2007) created long-lived, double-barred galaxy models in collisionless N -body simulations having dense inner disks, which they described as pseudo-bulges. They followed up with a more detailed study (Shen and Debattista 2009) that also made some predictions for observational tests. The secondary bars in their models indeed rotated at nonuniform rates, while their shape varied systematically with phase relative to that of the primary. These models prove that collisionless dynamics can support this behavior, but it is unclear that their initial conditions mimicked those that have given rise to double-barred galaxies in nature.

The possible consequence of gas inflow in these galaxies has attracted a lot of attention. Shlosman et al. (1989) speculated that bars within bars might lead to gas inflow over a wide dynamic range of scales, from global to the parsec scale where accretion onto a black hole might cause AGN activity. While inflows may have been observed (e.g., Haan et al. 2009), understanding of gas flow in these non-steady potentials remains rather preliminary (Maciejewski et al. 2002; Heller et al. 2007).

9.7 Buckling of Bars

Combes and Sanders (1981) first reported that the bars in their 3D simulations were thicker than the disk from which they had formed, and had acquired a pronounced “boxy” shape when viewed edge-on. Boxy isophotes in edge-on disk galaxies are now believed to be an indicator of a bar, as is supported by kinematic evidence in the gas (Merrifield and Kuijken 1999; Bureau and Athanassoula 2005).

The reason the bar thickened was explained by Raha et al. (1991), who showed that bars are subject to the buckling instability (► Sect. 8.1). The bar buckles because the formation of the bar created a structure supported by elongated orbits that stream along the bar in the near radial direction. Even though the ingoing and outgoing stars stream on different sides of the bar, the effective averaged σ_R has risen as a result without changing σ_z . The simulation by Raha et al. (1991) revealed that the buckling instability produced a large amplitude arch just before it saturated, after which the bar became thicker. The energy to increase vertical motion in the bar appeared to have been released by the further concentration of mass toward the bar center (see also Martinez-Valpuesta and Shlosman 2004). It is delightful that the evolution of one instability, the bar-forming mode, should create a new structure, the bar, that is itself unstable.

Bars still thicken in more recent simulations with grids having higher spatial resolution (e.g., ► Fig. 18-6), but do not seem to exhibit the spectacular arch reported by Raha et al. (1991) unless bi-symmetry is enforced. Low spatial resolution or significant gravity softening (which are equivalent) weakens the restoring force in (► 18.13) and artificially increases λ_J , which is the characteristic length for instability. Stronger gravity causes the preferred buckling modes to have shorter wavelength allowing, say, an upward arch on one side of the center and downward

arch on the other. Enforcing bi-symmetry prevents the bar from bending in this anti-symmetric manner, and forces it to buckle through the single arch mode.

9.8 Dynamical Friction on Bars


Friction between a rotating bar and a massive halo was first reported many years ago (Sellwood 1980), but the implications for dark matter halos have fueled a renewed intense study of the topic.


9.8.1 Theory

Dynamical friction (Chandrasekhar 1943) is the retarding force experienced by a massive perturber as it moves through a background sea of low-mass particles. It arises, even in a perfectly collisionless system, from the vector sum of the impulses the perturber receives from the particles as they are deflected by its gravitational field (see Appendix). Equivalently, friction can be viewed as the gravitational attraction on the perturber of the density excess, or wake, that develops behind it as it moves, as was nicely illustrated by Mulder (1983).

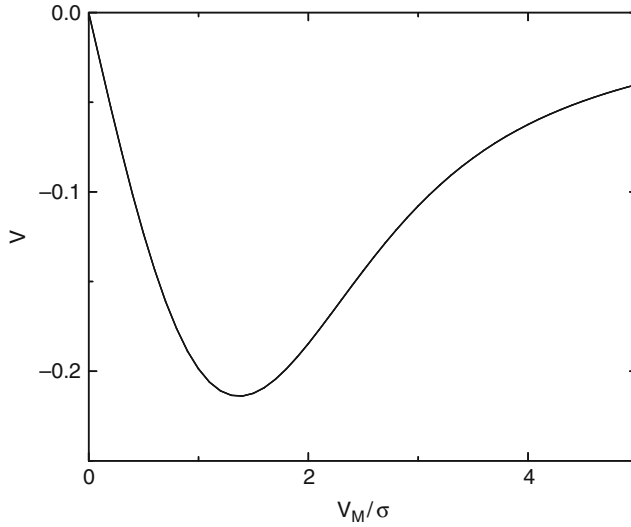
Chandrasekhar's formula (BT08 Eq. 8.6) for the acceleration of a perturber of mass M moving at speed v_M through a uniform background, density ρ , of noninteracting particles having an isotropic velocity distribution with a 1D rms velocity spread σ may be written as

$$\frac{dv_M}{dt} = 4\pi \ln \Lambda G^2 \frac{M\rho}{\sigma^2} V\left(\frac{v_M}{\sigma}\right). \quad (18.19)$$

The Coulomb logarithm is defined in the Appendix, and the dimensionless function V is drawn in  Fig. 18-15 for a Gaussian distribution of velocities; other velocity distributions would yield a different functional form. Physically, the retarding acceleration must vanish when the perturber is at rest and it must also tend to zero when the perturber moves so fast that the background particles receive only small impulses and the feeble wake lies far downstream from the perturber. Friction is strongest when the speed of the perturber is somewhat greater than the rms speeds of the background particles.

The simplifying assumptions in its derivation invalidate application of  to the physically more interesting problem of friction in a nonuniform medium in which the background particles are confined by a potential well and interact with the perturber repeatedly.

Repeated encounters between the perturber and the background particles require the more sophisticated treatment presented in Tremaine and Weinberg (1984b), who adopted a rotating potential perturbation in a gravitationally bound spherical halo of test particles. They showed, as did Lynden-Bell and Kalnajs (1972) for spiral waves, that lasting changes to the orbits of the halo particles appear to second order in the perturbing potential, and can occur only at resonances. They derived a daunting formula for the torque on the halo caused by the perturbation that sums the contributions from infinitely many resonances. The contribution at each resonance is proportional to the gradient of the DF near the phase-space location of the resonance, in a manner that is directly analogous to Landau damping. Weinberg (1985) computed the surprisingly large torque expected on a rotating bar, and his conclusion was confirmed in restricted tests (Little and Carlberg 1991; Hernquist and Weinberg 1992).



■ Fig. 18-15

The dimensionless acceleration function V defined in (18.19) for the case of a Gaussian distribution of velocities among the background particles. The function is negative because the acceleration is directed oppositely to the velocity

Weinberg and Katz (2007) pointed out that friction is dominated by a few important resonances. They estimated the widths of these resonances for a perturbation having constant pattern speed and finite amplitude, and argued that immense simulations would be needed to populate the resonance with sufficient particles to capture the correct net response. However, the loss of angular momentum from the perturber causes its pattern speed to change, and the resulting time dependence of the forcing frequency is a much more important factor in broadening the resonances; thus friction can in fact be captured correctly in simulations having moderate numbers of particles (Sellwood 2008a). Note that the pattern speed of an orbiting satellite rises as it loses angular momentum, while that of a bar usually decreases.

Despite the complicated language of resonant dynamics, the upshot is simply that the perturber induces a wake-like response in the halo, as was beautifully illustrated by Weinberg and Katz (2007, their Fig. 4). As for the infinite medium, friction can be thought of more simply as the torque between the perturber and the induced halo response. Sellwood (2006, his Fig. 2b) shows the lag angle between the forcing bar and the halo response, which is about 45° when friction is a maximum and gradually decreases to zero as the bar slows, until eventually friction ceases when the halo response corotates with the bar.

Lin and Tremaine (1983), for an orbiting satellite, and Sellwood (2006), for a rotating bar, demonstrated that the frictional drag on the perturbation scales with the mass of the perturber, M , the halo density, ρ , and the halo velocity dispersion, σ , exactly as in (18.19). Furthermore, the dimensionless function that describes the dependence on the angular speed of the perturber shares the general properties with $V(x)$ that it is negative (for reasonable non-rotating halos), and must $\rightarrow 0$ as $x \rightarrow \infty$, and that it should be $\propto x$ as $x \rightarrow 0$. Including self-gravity in the halo response causes a further slight increase in friction, but does not otherwise change the behavior.

9.8.2 Halo Density Constraint

Fully self-consistent simulations of bar formation in a live halo by Debattista and Sellwood (1998, 2000) showed that strong bars are indeed slowed rapidly. The fact that observed bars appear not to have been slowed (☛ Sect. 9.5) may imply an upper bound to the density of the dark matter halo in barred disk galaxies. Valenzuela and Klypin (2003) claimed a counter-example of a bar that does not experience much friction in a “cosmologically motivated” halo, even though their result disagreed with all others for strong bars (O’Neill and Dubinski 2003; Athanassoula 2003) and with theory!

Investigation of their anomalous result by Sellwood and Debattista (2006) revealed that friction can be avoided *temporarily* if the gradient of the DF at the most important resonance(s) has been flattened by earlier evolution, which they described as a metastable state. Lin and Tremaine (1983) reported similar behavior as a result of driving the perturber at constant frequency for a protracted period. They showed, as did Sellwood and Debattista (2006) and Villa-Vargas et al. (2009), that the full frictional drag resumes after some delay, the duration of which seems to vary stochastically (Sellwood and Debattista 2009). Delayed friction can happen only in simulations of disks in isolated, smooth halos, since any reasonable amount of halo substructure, or a tidal encounter, disturbs the delicate metastable state of the halo, causing friction to appear with its full force. Thus simulations that do not find strong friction from moderately dense halos (e.g., Klypin et al. 2009) have simply not been run for long enough.

While Debattista and Sellwood (2000) argued for near maximal disks, and their requirement for a low halo density is in agreement with the disk masses derived from fitting bar flow models (☛ Sect. 9.4), their constraint on the halo density may be specific to their adopted halo models. Thus additional careful studies of other halo models seem warranted.

9.8.3 Halo Density Reduction by Bars

While a full discussion of processes that may lower the dark matter density in the centers of halos is outside this review, a brief mention of the effect of bar friction is appropriate here.

Weinberg and Katz (2002) argued that the transfer of angular momentum from the bar to the halo could reduce the central density of the dark matter halo by a substantial factor. However, the possible density reduction is quite modest (Holley-Bockelmann et al. 2005; McMillan and Dehnen 2005; Sellwood 2008a) because the disk has only a finite amount angular momentum to give to the halo. Furthermore, as the disk loses angular momentum, its mass distribution contracts, and the deepening potential well further compresses the halo, which actually overwhelms the slight density reduction (Colín et al. 2006; Sellwood 2003).

10 Secular Evolution Within Disks

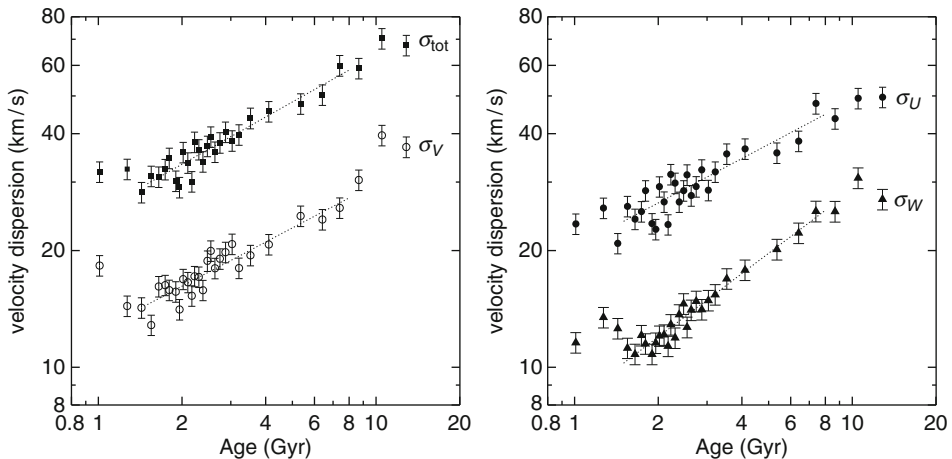
Most of the behavior discussed so far, such as the nonlinear evolution of instabilities, causes changes to the host galaxy on a dynamical timescale. The broad topic of secular evolution in galaxies describes changes that occur more gradually, such as the secular formation of pseudo-bulges (see Kormendy 1993; Kormendy and Kennicutt 2004, for excellent reviews), the

formation of rings (e.g., Buta and Combes 1996), or dynamical friction between components (► Sect. 9.8). The discussion in this section concerns processes that scatter disk stars only.

It has long been realized that old stars in the solar neighborhood have larger peculiar motions relative to the local standard of rest (hereafter LSR) than do young stars (e.g., Wielen 1977; Nordström et al. 2004; Aumer and Binney 2009). Postulating that older stars were born with larger random speeds, say in a thicker disk, is unattractive because it makes the present epoch of low velocity dispersion special. Some mechanism to scatter stars must therefore be invoked to create the larger random speeds of older stars.

The trends presented in ► Fig. 18-16 use the revised ages (Holmberg et al. 2007) assigned to the GCS stars (see Soderblom 2010, for a review). Figures 2, 3 and 4 of Aumer and Binney (2009) show the dispersions estimated as a function of color on the main sequence; while blue stars are necessarily young, red stars are expected to have a range of ages. The dispersions of the supposed oldest stars estimated by Holmberg et al. (2007) are no greater than those estimated by Aumer and Binney (2009) for their reddest stars, suggesting that the “older” bins in ► Fig. 18-16 include stars having a wide range of ages, as argued by Reid et al. (2007). It should be noted that the reported dispersions are simple second moments of the perhaps complex velocity distributions (see ► Fig. 18-1).

A reliable determination of the variation of dispersion as a function of time could provide another useful constraint on the scattering mechanism (e.g., BT08 ► Sect. 8.4). Quillen and Garnett (2000) and Seabroke and Gilmore (2007) argued that the dispersion may saturate for stars above a certain age; with a much older surge to account for the highest velocities. However, Aumer and Binney (2009) found better fits to the data with continuous acceleration, and deduced $\sigma \sim t^{0.35}$, with t being the current age of the stellar generation, in tolerable agreement, in fact, with 0.38 for the logarithmic slope of σ_{tot} in ► Fig. 18-16.



■ Fig. 18-16

The dispersion of stars in all three components, and the total dispersion of the GCS sample of solar neighborhood stars using revised ages (Holmberg et al. 2007). The radial, azimuthal, and vertical components are σ_U , σ_V , and σ_W , respectively, and $\sigma_{\text{tot}}^2 = \sigma_U^2 + \sigma_V^2 + \sigma_W^2$. The fitted straight lines do not include the first three or last two points

Three principal scattering agents have been discussed: dense gas clumps in the disk, massive black holes in the halo, and recurrent short-lived spirals. Note that the first two are essentially collisional processes that accelerate the relaxation rate (see the Appendix), while the changes caused by spirals can increase random motion without leading to a more relaxed DF. As none in isolation fits the data, a combination of spirals and gas clumps seems to be favored. Minor mergers and the effects of halo substructure are discussed in [Sect. 11](#).

An orbiting mass clump induces a collective spiral wake in the surrounding disk that enhances its mass and size by substantial factors (Julian and Toomre 1966), a complication that is ignored in many studies of cloud scattering. Since molecular gas is mostly concentrated in spiral arms (Nieten et al. 2006; Gratier et al. 2010; Efremov 2010), it is probably futile to draw a sharp distinction between spiral arms and the wakes of dense gas clumps, and a correct treatment would be to calculate the effects of the combined star and gas disk. Binney and Lacey (1988) took a step in this direction, but a full calculation may remain unreachable for some time if one tries to include a self-consistent treatment of the formation and dispersal of the gas clumps: molecular gas concentrations probably grow in the converging gas flow into a spiral arm, and are subsequently dispersed by star formation.

Treating spirals and mass clumps in the disk as distinct scattering agents may be justified, therefore, if the wakes of cloud complexes can be lumped with spirals into a single scattering agent that is distinct from the clouds that caused them. At the very least, this simplifying assumption separates the problem into tractable pieces.

10.1 Heating by Spirals

Lynden-Bell and Kalnajs (1972) showed that stars are scattered by a slowly changing potential perturbation only near resonances. More precisely, a spiral potential that grows and decays adiabatically, i.e., on a time-scale long compared with the orbital and epicyclic periods, will not cause a lasting change to a star's E and L_z . Wave-particle interactions become important near the resonances, where stars experience secular changes through “surfing” on the potential variations at CR or through a periodic forcing close to their epicyclic frequency at the Lindblad resonances. Either case produces a lasting change to a star's orbit.

The width of a resonance, i.e., the range of orbit frequencies of stars that are strongly affected, depends only on the amplitude of the potential when the perturbation is long-lived. But perturbations of shorter lifetimes have a broader range of frequencies and more stars experience lasting changes.

The discussion in [Sect. 2.4](#) and [Fig. 18-3](#) indicate that stars that lose (gain) L_z near the ILR (OLR) move onto more eccentric orbits, which is the root cause of heating by spirals. Exchanges at CR move stars to new orbits also, but with no change to the energy of noncircular motion, as discussed in [Sect. 10.2](#).

Significant heating by spiral waves over a large part of a disk requires them to be transient; a quasi-steady pattern, of the type envisaged by Bertin and Lin (1996) say, will cause localized heating at an exposed resonance, while stars elsewhere will move through the pattern without otherwise being affected. Barbanis and Woltjer (1967), Carlberg and Sellwood (1985), and Binney and Lacey (1988) calculated the heating caused by transient spirals. Jenkins and Binney (1990), De Simone et al. (2004), and Minchev and Quillen (2006) presented numerical studies of the consequences for a disk of test particles subject to some assumed set of spiral wave perturbations.

It is important to realize that the vertical oscillations of stars are little affected by spiral potential variations (► Sect. 2.6 and Carlberg 1987). In the absence of heavy clumps that can redirect disk velocities (⦿ Sect. 10.3), the increasing in-plane motions in simulations of initially cool, thin disks may ultimately cause the velocity ellipsoid to become sufficiently anisotropic as to cause it to thicken through mild buckling instabilities (⦿ Sect. 8.1). This may account for claims (e.g., McMillan and Dehnen 2007) that disks thicken due to spiral heating.

10.2 Churning by Transient Spirals

Studies of the metallicities and ages of nearby stars (Edvardsson et al. 1993; Nordström et al. 2004) found that older stars tend to have lower metallicities on average. As the ages of individual stars are disputed (Reid et al. 2007; Holmberg et al. 2007), the precise form of the relation is unclear. However, there seems to be general agreement that there is a spread of metallicities at each age, which is also supported by other studies (Chen et al. 2003; Haywood 2008; Stanghellini and Haywood 2010). The spread seems to be more than twice that expected from simple blurring of the gradient by stellar epicyclic excursions. In the absence of radial mixing, a metallicity spread amongst coeval stars is inconsistent with a simple chemical evolution model in which the metallicity of the disk rises monotonically in each annular bin.

Sellwood and Binney (2002) showed that scattering at CR causes very effective mixing. In a few Gyr, multiple transient spirals caused stars to diffuse in radius. Churning of the stellar disk occurs at corotation of the spirals with no associated heating and is able to account for the apparent metallicity spread with age. Roškar et al. (2008a, b) presented more detailed simulations that included infall, star formation, and feedback that confirmed this behavior. Schönrich and Binney (2009) developed the first chemical evolution model for the MW disk to include radial churning.

10.3 Cloud Scattering

Many years before the discovery of giant molecular gas clouds, Spitzer and Schwarzschild (1953) postulated their existence to account for the secular heating of disk stars. Lacey (1984) extended their calculation to 3D and concluded that cloud scattering should cause the vertical dispersion, σ_z , to be intermediate between the radial, σ_R , and azimuthal, σ_ϕ , components.¹⁶ This result seems physically plausible on energy equipartition grounds: scattering by massive clouds redirects the peculiar motions of stars through random angles, and therefore isotropizes the motions as far as the epicycle gyrations allow.

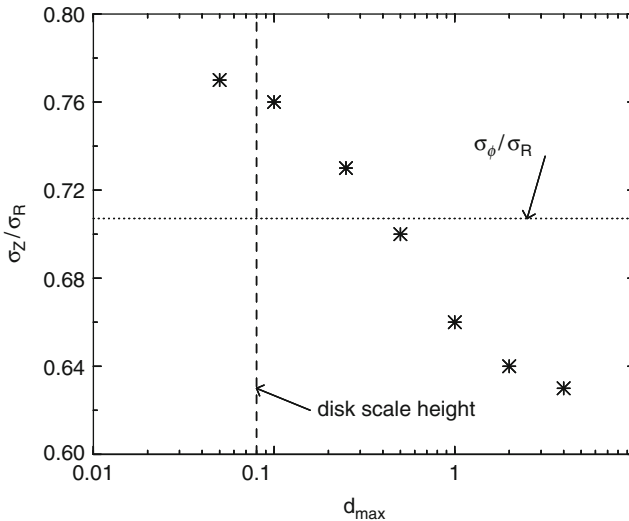
Despite the fact that redirecting peculiar motions happens much more rapidly than they can be increased by the same scatterers, the data do not reveal Lacey's expected axis ratio. The second moments of the velocity distribution of solar neighborhood stars in the three orthogonal directions (► Fig. 18-16 and Wielen 1977; Aumer and Binney 2009) satisfy the inequality $\sigma_z < \sigma_\phi < \sigma_R$. The ratio of the two in-plane components is in reasonable agreement with expectations from epicyclic motions, but the vertical component is the smallest, and this remains true for all groups when the stars are subdivided according to color or estimates of their ages. Gerssen et al. (2000) also observed a flattened ellipsoid in the disk of NGC 2895.

¹⁶The ratio $\sigma_R/\sigma_\phi \approx 2\Omega/\kappa$ (BT08, Eq. 8.117) is forced by epicyclic motions of disk stars.

Carlberg (1987) and Jenkins and Binney (1990) therefore developed the plausible argument that spirals drive up the in-plane components more rapidly than scattering is able to redirect those motions into the vertical direction, thereby accounting for the observed axis ratios of the velocity ellipsoid. Sellwood (2000) cited their argument as offering strong support for the transient spiral picture, but it now seems to be incorrect.

Ida et al. (1993) claimed that cloud scattering alone would lead to the vertical component being the smallest, with the precise axis ratio depending on the local slope of the rotation curve. Their simulations (Shiidsuke and Ida 1999), and others (e.g., Villumsen 1985; Hänninen and Flynn 2002), confirmed their expectation.

Sellwood (2008b) resolved this disagreement using simulations of test particles in the sheared sheet (see Sect. 3.6). Scattering by randomly distributed co-orbiting mass clumps confirmed the flattened velocity ellipsoid predicted by Ida et al. (1993). Figure 18-17 reveals why Ida's prediction differs from Lacey's: Lacey, as others (Spitzer and Schwarzschild 1953; Binney and Lacey 1988), assumed that cloud scattering is local, but the $\ln \Lambda$ term in the formulae in the Appendix implies that distant encounters dominate any scattering process in 3D, at least to distances a few times the disk thickness. Distant scatterers in the flattened geometry of a disk must predominantly affect the in-plane star velocities, and couple much less strongly to the vertical component. Figure 18-17 shows the equilibrium ratio σ_z/σ_R when scatterers beyond the finite range d_{\max} artificially exert no forces. The ratio settles to something close to Lacey's energy equipartition prediction when none but the closest heavy clumps perturb the stars, but the equilibrium ellipsoid flattens in separate experiments as more distant clouds are included, tending toward Ida's result with no artificial cutoff.



■ Fig. 18-17

The equilibrium axis ratio of the velocity ellipsoid of particles plotted as a function of the limiting range of the perturbation forces from the heavy particles. See Sellwood (2008b) for a description of the calculations

Thus the *shape* of the local velocity ellipsoid, [▶ Fig. 18-16](#), cloud scattering and does not, as seemed attractive, require spiral arm scattering. However, the data do not imply that spirals are unimportant: cloud scattering seems unable to generate the random speeds of the oldest stars (e.g., Lacey 1991; Hänninen and Flynn 2002), and there are hints in [▶ Fig. 18-16](#) of some evolution of the velocity ellipsoid shape that may demand a compound origin.

10.4 Black Holes in the Halo

The possibility that the dark matter halos of galaxies are made up of massive black holes (BHs) is not yet excluded. Lacey and Ostriker (1985) calculated the consequences for the stars in the disk of the MW, assuming the BHs to have orbits characteristic of a pressure-supported halo and to impart impulses to disk stars as they pass through the disk. The high speeds of their encounters with disk stars would cause the velocity dispersion to rise as $t^{1/2}$, while the predicted shape of the velocity ellipsoid is in reasonable agreement with that observed.

Lacey and Ostriker (1985) addressed a number of issues with their model, such as the X-ray accretion luminosity as the BHs pass through the gas disk, the accumulation of BHs in the galactic center through dynamical friction, and the survival of dwarf galaxies. They also acknowledged that it does not predict the correct variation of σ_{tot} with Galactic radius. BT08 (Sect. 7.4.4) added that the idea could be ruled out on the grounds that wide binary star systems would be disrupted too quickly.

If disk scattering is dominated by spirals, as argued below, then scattering by BHs would be needed merely to redirect the peculiar motions into the vertical direction. This reviewer is not aware of any such calculation, but since passing BHs scatter stars in the direction perpendicular to their orbits, it seems unlikely that they could redirect peculiar motion without also increasing in-plane motions. However, if this expectation is too pessimistic and the desired axis ratio could be achieved with lower BH masses, many of the other objections are weakened.

10.5 Discussion

The extraordinary phase-space structure of the solar neighborhood (Dehnen 1998; Nordström et al. 2004, and [▶ Fig. 18-1](#)) indicates that there is little in the way of an underlying smooth component and the stellar distribution is broken into several “streams” (Bovy et al. 2009). The features are too substantial to have simply arisen from groups of stars that were born with similar kinematics (e.g., Eggen 1996), as confirmed in detailed studies (Famaey et al. 2007; Bensby et al. 2007; Bovy and Hogg 2009).

Thus it is clear that the entire DF has been sculptured by dynamical processes. Were the large part of the spread in velocities caused by multiple scatterings off molecular clouds in the disk, or off black holes in the halo, the distribution should approximate the simple double Gaussian proposed by K. Schwarzschild (see BT08 [▶ Sect. 4.4.3](#)). The vertical velocity distribution (Nordström et al. 2004), on the other hand, does have a relaxed appearance, as noted by Seabroke and Gilmore (2007).

Various dynamical agents have been proposed to account for kinematic features in the solar neighborhood. Kalnajs (1991) argued that the OLR of the bar in the MW might be close to the solar circle. Features in the subsequently released HIPPARCOS data have also been attributed

to the OLR of the bar (Raboud et al. 1998; Dehnen 2000; Fux 2001), while Sellwood (2010) attributed another to a recent ILR.

De Simone et al. (2004) were able to produce distributions of stars with a similar degree of substructure in simulations of test particles moving in the adopted potential perturbations representing a succession of short-lived spiral transients; see also Minchev and Quillen (2006). Other models that included both bars and spirals were presented by Quillen (2003), Chakrabarty (2007), and Antoja et al. (2009), while Helmi et al. (2006) suggested that some of the substructure may also be caused by satellite infall (see also Sect. 11).

Since spirals (and bars) are inefficient at exciting vertical motions, another scattering process needs to be invoked to redirect in-plane random motion into vertical motion. The influence of clumps in the disk, which are known to exist, seems consistent with the relaxed appearance of the vertical velocity distribution. Furthermore, the observed axes of the velocity ellipsoid, except possibly for the youngest stars (Fig. 18-16), are consistent with the prediction of cloud scattering. Thus cloud scattering seems to be just sufficient to redirect velocities, which they are good at, but not to contribute significantly to heating.

11 Fragility of Disks

Most of the stars in spiral galaxies reside in remarkably thin disks. In a photometric study of edge-on galaxies, Yoachim and Dalcanton (2006) found that the fraction of the baryons in the thin disk component is in the range 70–90%, with higher fractions being characteristic of more massive galaxies. The so-called superthin disks are even more extreme; the disk of the low-luminosity galaxy UGC 7321, e.g., has a radial scale-length some 14 times that of its characteristic thickness (Matthews 2000).

Hierarchical galaxy formation scenarios predict that galaxy formation is far from monolithic, with occasional major mergers with other halos, and frequent minor mergers. Stellar disks that have formed in the centers of the halos are torn apart in major mergers (Barnes and Hernquist 1992), but the consequences of minor mergers are harder to determine from simulations (e.g., Walker et al. 1996). Tóth and Ostriker (1992) argued that the existence of thin disks in galaxies can be used to constrain the rate of minor mergers. Their numerical estimates have been criticized on various grounds (e.g., Huang and Carlberg 1997; Sellwood et al. 1998; Velázquez and White 1999), but it is clear that a tight constraint remains.

Wyse (this volume) stresses that the thick disk of the MW contains only very old stars, and that the ages of thin disk stars stretch back 10 Gyr. This fact would seem to imply that the last galactic merger to have stirred the MW disk occurred some 10 Gyr ago, and that no comparable disturbance could have occurred since. The MW may not be unique in this respect, as Mould (2005) found that thick disks in four nearby galaxies also appear to be old.

Stewart et al. (2008) estimated the rate of mergers in the current Λ CDM cosmology. They concluded that 95% of parent halos of some $10^{12} h^{-1} M_{\odot}$ will have merged with another halo of at least 20% of its mass in the last 10 Gyr, and consequently the disk hosted by the parent must somehow survive in most cases.

Possibly only a small fraction of infalling satellites pose a real threat to a disk. All satellites will be tidally stripped as they fall into the main halo, and some may even dissolve completely before they can affect the disk – the Sagittarius dwarf galaxy appears to be a good example (Law

et al. 2005). A massive satellite loses orbital energy through dynamical friction (☛ Sect. 9.8) causing it to settle deeper into the main halo. Its survival depends on its density (see BT08☛ Sect. 8.3); it will be stripped of its loosely bound envelope until its mean density, $\bar{\rho}$, is about one third the mean density of the halo interior to its orbital radius (BT08 Eq. 8.92), and disrupted entirely as it reaches a radius in the main halo where even its central density falls below the mean interior halo density. This process has been studied in more detail by Boylan-Kolchin and Ma (2007) and by Choi et al. (2009). Thus the vulnerability of disks depends on the inner densities, in both dark matter and baryons, of the accreted sub-halos.

If even a single, moderately massive core survives to the inner halo, it could cause an unacceptable increase in the disk thickness. Vertical heating of the disk occurs when a passing or penetrating satellite is able to increase the vertical motions of the disk stars. High-speed passages therefore deposit little energy into disk motions, but if the satellite's orbit remains close to the disk mid-plane, then its vertical frequency will couple strongly to that of some of the disk stars and heating will be rapid. Indeed, Read et al. (2008) argued that the accretion event(s) that created the thick disk of the MW most probably resulted from the infall of a sub-halo whose orbit plane was inclined at $10\text{--}20^\circ$ to that of the disk. The energy deposited could take the form of exciting bending waves in the disk that can propagate radially until they are damped at vertical resonances (Sellwood et al. 1998).

Kazantzidis et al. (2009) simulated several minor mergers with sub-halos in detailed models with plausible parameters. They found that the disk is substantially thickened and heated by the mergers, although they noted that their simulations lacked a gaseous component. Simulations that include a gas component cannot resolve the small-scale physical processes of gas fragmentation, star formation, feedback, etc., and therefore the behavior of the dissipative gas component necessarily includes somewhat ad hoc prescriptions for these aspects of sub-grid physics. Hopkins et al. (2009) stressed that gas in mergers settles quickly to begin to form a new disk; however, it is the fate of the stars that had been formed prior to the merger that is the principal concern. Kazantzidis et al. (2009) argued it is possible that a dissipative component in the disk could absorb some of the orbital energy of the satellite which would reduce the heating of the stellar disk, and such an effect appears to have occurred in the simulations by Moster et al. (2010). A clear conclusion has yet to emerge from this ongoing research effort, but exclusively old, thick disks together with the prevalence of thin disks pose a substantial, though perhaps surmountable, challenge to the Λ CDM model.

It should be noted that ideas to reduce the density of the inner main halo through frictional energy loss to halo (e.g., El-Zant et al. 2001; Mashchenko et al. 2006; Romano-Díaz et al. 2008a) require the kinds of dense massive fragments that are themselves a danger to the disk. This will be of less importance in the early stages of galaxy assembly, but disk survival adds the requirement that any such process be completed quickly.

12 Conclusions

This review has focused on understanding the mechanisms that underlie the various instabilities and processes that affect the structure of galaxy disks. The discussion has referred extensively to simulations that are sufficiently simple that they clearly illustrate each particular aspect of the behavior. A complementary, and powerful, approach is to add additional physical processes to simulations, with the aim of improving realism to address otherwise inaccessible questions.

However, the increased complexity of the behavior that comes with increasing realism makes a deep understanding of the results harder to achieve. Hopefully, the behavior described here will form a solid basis on which to build as the challenges presented by more realistic systems are addressed.

A contrast with accretion disk theory seems appropriate here. Shakura and Sunyaev (1973) proposed a scaling relation for turbulent viscosity that led to rapid progress in modeling accretion disks some 25 years before the likely origin of the viscosity was identified (Balbus and Hawley 1998). A theory for the structural evolution of galaxies seems much farther away. While most of the important physical processes may have been identified, exactly how they drive evolution is still not fully understood, and even the evolutionary path remains vague. In the absence of this understanding, simplifying scaling laws cannot be identified with confidence, and it seems best to work at improved understanding of the mechanisms at play.

Galaxy dynamics has made immense strides in the 45 years since the chapter by Oort (1965) in the corresponding volume of the preceding series. Understanding of local wave mechanics (▶ Sect. 3) and the mechanisms for the principal global gravitational instabilities (▶ Sects. 4 and 6) is well advanced. Bending waves and global buckling modes (▶ Sect. 8) are mostly understood, but not at quite the same level, while lop-sided modes (▶ Sect. 5) perhaps still require a some more concerted effort. Understanding of bar structure (▶ Sect. 9) and the role of bars in galaxy evolution has developed beyond recognition, and disk heating (▶ Sect. 10) seems more solidly understood, with our confidence being strongly boosted by the extremely valuable data from Nordström et al. (2004).

But answers to a number of major questions of galaxy dynamics are still incomplete. Although the bar-forming instability is now understood (▶ Sect. 4.1), it does not provide a clear picture of why only a little over half of all bright disk galaxies are barred (▶ Sect. 9.2). After bars, spirals arms are the second most prominent feature of disk galaxies and are probably responsible for the most dynamical evolution (▶ Sect. 10), yet a deep understanding of their origin (▶ Sect. 7) remains elusive. The good progress made in recent years to understand the warps of galaxy disks (▶ Sect. 8) still has not supplied a satisfactory account of their incidence.

Many of these outstanding issues may be bound up with how galaxies form, our understanding of which is currently making particularly rapid progress.

13 Appendix

A13.1 Relaxation Time in Spheroids and Disks

A test particle moving at velocity \mathbf{v} along a trajectory that passes a stationary field star of mass m with impact parameter b is deflected by the attraction of the field star. For a distant passage, it acquires a transverse velocity component $|\mathbf{v}_\perp| \simeq 2Gm/(bv)$ to first order (BT08 Eq. 1.30). Encounters at impact parameters small enough to produce deflections where this approximation fails badly are negligibly rare and relaxation is driven by the cumulative effect of many small deflections.

If the density of field stars is n per unit volume, the test particle will encounter $\delta n = 2\pi b \delta b n v$ stars per unit time with impact parameters between b and $b + \delta b$. Assuming stars to have equal masses, each encounter at this impact parameter produces a randomly directed \mathbf{v}_\perp that will cause a mean square net deflection per unit time of

$$\delta v_\perp^2 \simeq \left(\frac{2Gm}{bv} \right)^2 \times 2\pi b \delta b n v = \frac{8\pi G^2 m^2 n}{v} \frac{\delta b}{b}. \quad (\text{A1})$$

The total rate of deflection from all encounters is the integral over impact parameters, yielding

$$v_{\perp}^2 = \frac{8\pi G^2 m^2 n}{v} \int_{b_{\min}}^{b_{\max}} \frac{db}{b} = \frac{8\pi G^2 m^2 n}{v} \ln \Lambda, \quad (\text{A2})$$

where $\ln \Lambda \equiv \ln(b_{\max}/b_{\min})$ is the Coulomb logarithm. Typically one chooses the lower limit to be the impact parameter of a close encounter, $b_{\min} \simeq 2Gm/v^2$, for which $|v_{\perp}|$ is overestimated by the linear formula, while the upper limit is, say, the half-mass (or effective) radius, R , of the stellar distribution beyond which the density decreases rapidly. The vagueness of these definitions is not of great significance to an estimate of the overall rate because we need only the logarithm of their ratio. The Coulomb logarithm implies equal contributions to the integrated deflection rate from every decade in impact parameter simply because the diminishing gravitational influence of more distant stars is exactly balanced by their increasing numbers.

Note that the first order deflections that give rise to this steadily increasing random energy come at the expense of second order reductions in the forward motion of the same particles that we have neglected (Hénon 1973). Thus the system does indeed conserve energy, as it must.

We define the **relaxation time** to be the time needed for $v_{\perp}^2 \simeq v^2$, where v is the typical velocity of a star. Thus

$$\tau_{\text{relax}} = \frac{v^3}{8\pi G^2 m^2 n \ln \Lambda}. \quad (\text{A3})$$

To order of magnitude, a typical velocity $v^2 \approx GNm/R$, where N is the number of stars each of mean mass m , yielding $\Lambda \approx N$. Defining the dynamical time to be $\tau_{\text{dyn}} = R/v$ and setting $N \sim R^3 n$, we have

$$\tau_{\text{relax}} \approx \frac{N}{6 \ln N} \tau_{\text{dyn}}, \quad (\text{A4})$$

which shows that the collisionless approximation is well satisfied in galaxies, which have $10^8 \lesssim N \lesssim 10^{11}$ stars. Including the effect of a smooth dark matter component in this estimate would increase the typical velocity, v , thereby further lengthening the relaxation time.

This standard argument, however, assumed a pressure-supported quasi-spherical system in several places. Rybicki (1972) pointed out that the flattened geometry and organized streaming motion within disks affects the relaxation time in two important ways. First, the assumption that the typical encounter velocity is comparable to the orbital speed $v = (GNm/R)^{1/2}$ is clearly wrong; stars move past each other at the typical random speeds in the disk, say βv with $\beta \sim 0.1$, causing larger deflections and decreasing the relaxation time by a factor β^3 .

Second, the distribution of scatterers is not uniform in 3D, as was implicitly assumed in (A1). Assuming a razor thin disk, changes the volume element from $2\pi v b \delta b$ for 3D to $2v \delta b$ in 2D, which changes the integrand in (A2) to b^{-2} and replaces the Coulomb logarithm by the factor $(b_{\min}^{-1} - b_{\max}^{-1})$. In 2D therefore, relaxation is dominated by close encounters. Real galaxy disks are neither razor thin, nor spherical, so the spherical dependence applies at ranges up to the typical disk thickness, z_0 , beyond which the density of stars drops too quickly to make a significant further contribution to the relaxation rate. Thus we should use $\Lambda \simeq z_0/b_{\min}$ for disks. More significantly, the local mass density is also higher, so that $N \sim R^2 z_0 n$. These considerations shorten the relaxation time by the factor $(z_0/R) \ln(R/z_0)$. An additional effect of flattened distribution of scatterers is to determine the shape of the equilibrium velocity ellipsoid, as discussed in Sect. 10.3.

A third consideration for disks is that the mass distribution is much less smooth than is the case in the bulk of pressure supported galaxies. A galaxy disk generally contains massive

star clusters and giant molecular clouds whose influence on the relaxation rate turns out to be non-negligible (see [Sect. 10](#)).

Acknowledgments

The author is indebted to James Binney, Victor Debattista, Agris Kalnajs, Juntai Shen, Alar Toomre, and Scott Tremaine for numerous valuable comments on a draft of this chapter.

Cross-References

- ▶ [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- ▶ [History of DarkMatter in Galaxies](#)
- ▶ [Mass Distribution and Rotation Curve in the Galaxy](#)

References

- Agertz, O., Teyssier, R., & Moore, B. 2010, arXiv:1004.0005
- Aguerre, J. A. L., Méndez-Abreu, J., & Corsini, E. M. 2009, *A&A*, 495, 491
- Antoja, T., Valenzuela, O., Pichardo, B., Moreno, E., Figueras, F., & Fernández, D. 2009, *ApJL*, 700, L78
- Aoki, S., Noguchi, M., & Iye, M. 1979, *PASJ*, 31, 737
- Araki, S. 1985, PhD thesis, MIT
- Araki, S. 1987, *AJ*, 94, 99
- Athanassoula, E. 1992, *MNRAS*, 259, 345
- Athanassoula, E. 2002, *ApJL*, 569, L83
- Athanassoula, E. 2003, *MNRAS*, 341, 1179
- Athanassoula, E., Bosma, A., & Papaioannou, S. 1987, *A&A*, 179, 23
- Athanassoula, E., Lambert, J. C., & Dehnen, W. 2005, *MNRAS*, 363, 496
- Athanassoula, E., Romero-Gómez, M., & Masdemont, J. J. 2009, *MNRAS*, 394, 67
- Athanassoula, E., & Sellwood, J. A. 1986, *MNRAS*, 221, 213
- Aumer, M., & Binney, J. J. 2009, *MNRAS*, 397, 1286
- Bailin, J. 2003, *ApJL*, 583, L79
- Balbus, S. A., & Hawley, J. F. 1998, *Rev Mod Phys*, 70, 1
- Baldwin, J. E., Lynden-Bell, D., & Sancisi, R. 1980, *MNRAS*, 193, 313
- Barazza, F. D., Jogee, S., & Marinova, I. 2008, *ApJ*, 675, 194
- Barazza, F. D. et al. 2009, *A&A*, 497, 713
- Barbanis, B., & Woltjer, L. 1967, *ApJ*, 150, 461
- Barnes, J. E., & Hernquist, L. 1992, *ARAA*, 30, 705
- Benedict, G. F., Howell, A., Jorgensen, I., Kenney, J., & Smith, B. J. 2002, *AJ*, 123, 1411
- Bensby, T., Oey, M. S., Feltzing, S., & Gustafsson, B. 2007, *ApJL*, 655, L89
- Berentzen, I., Athanassoula, E., Heller, C. H., & Fricke, K. J. 2004, *MNRAS*, 347, 220
- Berentzen, I., Shlosman, I., Martínez-Valpuesta, I., & Heller, C. H. 2007, *ApJ*, 666, 189
- Bertin, G., & Lin, C. C. 1996, *Spiral Structure in Galaxies* (Cambridge, MA: MIT)
- Binney, J., Jiang, I., & Dutta, S. 1998, *MNRAS*, 297, 1237
- Binney, J. J., & Lacey, C. G. 1988, *MNRAS*, 230, 597
- Binney, J., & Tremaine, S. 2008, *Galactic Dynamics* (2nd ed.; Princeton: Princeton University Press) (BT08)
- Block, D. L., Freeman, K. C., Jarrett, T. H., Puerari, I., Worthey, G., Combes, F., & Groess, R. 2004, *A&A*, 425, L37
- Bosma, A. 1991, in *Warped Disks and Inclined Rings Around Galaxies*, eds. S. Casertano, P. D. Sackett, & F. H. Briggs (Cambridge: Cambridge University Press), 181
- Bosma, A. 1996, in *IAU Colloq. 157, Barred Galaxies*, eds. R. Buta, D. A. Crocker, & B. G. Elmegreen (San Francisco: ASP Conf series 91), 132
- Bottema, R. 1996, *A&A*, 306, 345
- Bournaud, F., Combes, F., & Semelin, B. 2005, *MNRAS*, 364, L18
- Bovy, J., & Hogg, D. W. 2009, arXiv:0912.3262
- Bovy, J., Hogg, D. W., & Roweis, S. T. 2009, *ApJ*, 700, 1794

- Boylan-Kolchin, M., & Ma, C.-P. 2007, *MNRAS*, 374, 1227
- Briggs, F. H. 1990, *ApJ*, 352, 15
- Bureau, M., & Athanassoula, E. 2005, *ApJ*, 626, 159
- Buta, R., & Combes, F. 1996, *Fund Cosmic Phys*, 17, 95
- Buta, R. J., Knapen, J. H., Elmegreen, B. G., Salo, H., Laurikainen, E., Elmegreen, D. M., Puerari, I., & Block, D. L. 2009, *AJ*, 137, 4487
- Camm, G. L. 1950, *MNRAS*, 110, 305
- Carlberg, R. G. 1987, *ApJ*, 322, 59
- Carlberg, R. G., & Freedman, W. L. 1985, *ApJ*, 298, 486
- Carlberg, R. G., & Sellwood, J. A. 1985, *ApJ*, 292, 79
- Carollo, C. M., Stiavelli, M., & Mack, J. 1998, *AJ*, 116, 68
- Chakrabarty, D. 2007, *A&A*, 467, 145
- Chandrasekhar, S. 1943, *ApJ*, 97, 255
- Chemin, L., & Hernandez, O. 2009, *A&A*, 499, L25
- Chen, L., Hou, J. L., & Wang, J. J. 2003, *AJ*, 125, 1397
- Chirikov, B. V. 1979, *Phys Rep*, 52, 265–379
- Choi, J.-H., Weinberg, M. D., & Katz, N. 2009, *MNRAS*, 400, 1247
- Christodoulou, D. M., Shlosman, I., & Tohline, J. E. 1995, *ApJ*, 443, 551
- Colín, P., Valenzuela, O., & Klypin, A. 2006, *ApJ*, 644, 687
- Combes, F., & Sanders, R. H. 1981, *A&A*, 96, 164
- Contopoulos, G. 1980, *A&A*, 81, 198
- Corbelli, E., & Walterbos, R. A. M. 2007, *ApJ*, 669, 315
- Corsini, E. M. 2008, in *Formation and Evolution of Galaxy Bulges*, IAU Symp. 245 (Dordrecht: Kluwer), 125
- Corsini, E. M., Debattista, V. P., & Aguerri, J. A. L. 2003, *ApJL*, 599, L29
- Courteau, S., Andersen, D. R., Bershady, M. A., MacArthur, L. A., & Rix, H.-W. 2003, *ApJ*, 594, 208
- Cox, A. L., Sparke, L. S., van Moorsel, G., & Shaw, M. 1996, *AJ*, 111, 1505
- Curir, A., Mazzei, P., & Murante, G. 2006, *A&A*, 447, 453
- Cuzzi, J. N., et al. 2010, *Science*, 327, 1470
- Davoust, E., & Contini, T. 2004, *A&A*, 416, 515
- Debattista, V. P., & Sellwood, J. A. 1998, *ApJL*, 493, L5
- Debattista, V. P., & Sellwood, J. A. 1999, *ApJL*, 513, L107
- Debattista, V. P., & Sellwood, J. A. 2000, *ApJ*, 543, 704
- Debattista, V. P., & Shen, J. A. 2007, *ApJL*, 654, L127
- Dehnen, W. 1998, *AJ*, 115, 2384
- Dehnen, W. 2000, *AJ*, 119, 800
- Dekel, A., & Shlosman, I. 1983, in *IAU Symposium 100, Internal Kinematics and Dynamics of Galaxies*, ed. E. Athanassoula (Dordrecht: Reidel), 187
- De Simone, R. S., Wu, X., & Tremaine, S. 2004, *MNRAS*, 350, 627
- Dobbs, C. L., Theis, C., Pringle, J. E., & Bate, M. R. 2010, *MNRAS*, (in press)
- Donner, K. J., & Thomasson, M. 1994, *A&A*, 290, 475
- Dubinski, J., Berentzen, I., & Shlosman, I. 2009, *ApJ*, 697, 293
- Dubinski, J., & Chakrabarty, D. 2009, *ApJ*, 703, 2068
- Dubinski, J., Gauthier, J.-R., Widrow, L., & Nickerson, S. 2008, in *Formation and Evolution of Galaxy Disks*, ed. J. G. Funes SJ & E. M. Corsini (San Francisco: ASP 396), 321
- Dubinski, J., & Kuijken, K. 1995, *ApJ*, 442, 492
- Dury, V., de Rijcke, S., Debattista, V. P., & Dejonghe, H. 2008, *MNRAS*, 387, 2
- Earn, D. J. D., & Lynden-Bell, D. 1996, *MNRAS*, 278, 395
- Edvardsson, B., Andersen, B., Gustafsson, B., Lambert, D. L., Nissen, P. E., & Tomkin, J. 1993, *A&A*, 275, 101
- Efremov, Yu. N. 2010, *MNRAS*, to appear (arXiv:1002.4555)
- Efstathiou, G., Lake, G., & Negroponte, J. 1982, *MNRAS*, 199, 1069
- Eggen, O. J. 1996, *AJ*, 112, 1595
- Elmegreen, B. 1996, in *IAU Colloq. 157, Barred Galaxies*, eds. R. Buta, D. A. Crocker, & B. G. Elmegreen (San Francisco: ASP Conf series 91), 197
- Elmegreen, B. G., & Thomasson, M. 1993, *A&A*, 272, 37
- Elmegreen, B. G., Elmegreen, D. M., Knapen, J. H., Buta, R. J., Block, D. L., & Puerari, I. 2007, *ApJL*, 670, L97
- Elmegreen, D. M., Elmegreen, B. G., & Bellin, A. D. 1990, *ApJ*, 364, 415
- El-Zant, A., Shlosman, I., & Hoffman, Y. 2001, *ApJ*, 560, 636
- Englmaier, P., & Gerhard, O. 1997, *MNRAS*, 287, 57
- Englmaier, P., & Shlosman, I. 2004, *ApJL*, 617, L115
- Erwin, P. 2005, *MNRAS*, 364, 283
- Erwin, P., & Sparke, L. S. 2002, *AJ*, 124, 65
- Eskridge, P. B., et al. 2000, *AJ*, 119, 536
- Evans, N. W., & Read, J. C. A. 1998, *MNRAS*, 300, 106
- Famaey, B., Pont, F., Luri, X., Udry, S., Mayor, M., & Jorissen, A. 2007, *A&A*, 461, 957
- Fathi, K., Beckman, J. E., Piñol-Ferrer, N., Hernandez, O., Martínez-Valpuesta, I., & Carignan, C. 2009, *ApJ*, 704, 1657
- Fridman, A. M., & Polyachenko, V. L. 1984. *Physics of Gravitating Systems* (New York: Springer)
- Friedli, D., & Martinet, L. 1993, *A&A*, 277, 27

- Fuchs, B., Dettbarn, C., & Tsuchiya, T. 2005, *A&A*, 444, 1
- Fux, R. 1999, *A&A*, 345, 787
- Fux, R. 2001, *A&A*, 373, 511
- García-Ruiz, I., Kuijken, K., & Dubinski, J. 2002a, *MNRAS*, 337, 459
- García-Ruiz, I., Sancisi, R., & Kuijken, K. 2002b, *A&A*, 394, 769
- Gerssen, J., Kuijken, K., & Merrifield, M. R. 2000, *MNRAS*, 317, 545
- Goldreich, P., & Lynden-Bell, D. 1965a, *MNRAS*, 130, 97
- Goldreich, P., & Lynden-Bell, D. 1965b, *MNRAS*, 130, 125
- Goldreich, P., & Tremaine, S. 1978, *ApJ*, 222, 850
- Gratier, P. et al. 2010, *A&A*, to appear (arXiv:1003.3222)
- Grosbøl, P., Patsis, P. A., & Pompei, E. 2004, *A&A*, 423, 849
- Haan, S., Schinnerer, E., Emsellem, E., Garca-Burillo, S., Combes, F., Mundell, C. G., & Rix, H.-W. 2009, *ApJ*, 692, 1623
- Hänninen, J., & Flynn, C. 2002, *MNRAS*, 337, 731
- Haywood, M. 2008, *MNRAS*, 388, 1175
- Heller, C. H., Shlosman, I., & Athanassoula, E. 2007, *ApJ*, 657, L65
- Heller, C., Shlosman, I., & Englmaier, P. 2001, *ApJ*, 553, 661
- Helmi, A., Navarro, J. F., Nordström, B., Holmberg, J., Abadi, M. G., & Steinmetz, M. 2006, *MNRAS*, 365, 1309
- Hénon, M. 1973, in *Dynamical Structure and Evolution of Stellar Systems*, eds. L. Martinet, & M. Mayor (Sauverny: Geneva Observatory), 182
- Hernquist, L., & Weinberg, M. D. 1992, *ApJ*, 400, 80
- Hill, G. W. 1878 *Am. J. Math.*, 1, 5
- Hockney, R. W., & Brownrigg, D. R. K. 1974, *MNRAS*, 167, 351
- Hohl, F. 1971, *ApJ*, 168, 343
- Holley-Bockelmann, K., Weinberg, M., & Katz, N. 2005, *MNRAS*, 363, 991
- Holmberg, J., & Flynn, C. 2004, *MNRAS*, 352, 440
- Holmberg, J., Nordström, B., & Andersen, J. 2007, *A&A*, 475, 519
- Holmberg, J., Nordström, B., & Andersen, J. 2009, *A&A*, 501, 941
- Hopkins, P. F., Cox, T. J., Younger, J. D., & Hernquist, L. 2009, *ApJ*, 691, 1168
- Huang, S., & Carlberg, R. G. 1997, *ApJ*, 480, 503
- Hunter, C., & Toomre, A. 1969, *ApJ*, 155, 747
- Ida, S., Kokuba, E., & Makino, J. 1993, *MNRAS*, 263, 875
- Ideta, M. 2002, *ApJ*, 568, 190
- Jalali, M. A. 2007, *ApJ*, 669, 218
- James, R. A., & Sellwood, J. A. 1978, *MNRAS*, 182, 331
- Jeans, J. H. 1923, *MNRAS*, 84, 60
- Jeans, J. H. 1929, *Astronomy and Cosmogony* (Cambridge: Cambridge University Press)
- Jenkins, A., & Binney, J. J. 1990, *MNRAS*, 245, 305
- Jiang, I., & Binney, J. 1999, *MNRAS*, 303, L7
- Jog, C. J., & Combes, F. 2009, *Phys Rep*, 471, 75
- Jog, C. J., & Solomon, P. M. 1992, *ApJ*, 387, 152
- Jogee, S. et al. 2004, *ApJL*, 615, L105
- Julian, W. H., & Toomre, A. 1966, *ApJ*, 146, 810
- Kahn, F. D., & Woltjer, L. 1959, *ApJ*, 130, 705
- Kalnajs, A. J. 1965, PhD thesis, Harvard University
- Kalnajs, A. J. 1972, *ApJ*, 175, 63
- Kalnajs, A. J. 1976, *ApJ*, 205, 751
- Kalnajs, A. J. 1978, in *IAU Symp. 77, Structure and Properties of Nearby Galaxies* eds. E. M. Berkhuisjen & R. Wielebinski (Dordrecht: Reidel), 113
- Kalnajs, A. J. 1991, in *Dynamics of Disc Galaxies*, ed. B. Sundelius (Gothenburg: Göteborgs University), 323
- Kazantzidis, S., Zentner, A. R., Kravtsov, A. V., Bullock, J. S., & Debattista, V. P. 2009, *ApJ*, 700, 1896
- Khoperskov, A. V., Just, A., Korchagin, V. I., & Jalali, M. A. 2007, *A&A*, 473, 31
- Klypin, A., Valenzuela, O., Colin, P., & Quinn, T. 2009, *MNRAS*, 398, 1027
- Korchagin, V., Orlova, N., Kikuchi, N., Miyama, S. M., & Moiseev, A. V. 2005, arXiv:astro-ph/0509708
- Kornreich, D. A., Lovelace, R. V. E., & Haynes, M. P. 2002, *ApJ*, 580, 705
- Kormendy, J. 1993, in *IAU Symp. 153, Galactic Bulges*, eds. H. Dejonghe & H. Habing (Dordrecht: Kluwer), 209
- Kormendy, J., & Kennicutt, R. C. 2004, *ARAA*, 42, 603
- Kormendy, J., & Norman, C. A. 1979, *ApJ*, 233, 539
- Kranz, T., Slyz, A. D., & Rix, H.-W. 2003, *ApJ*, 586, 143
- Kuijken, K. 1991, *ApJ*, 376, 467
- Kulsrud, R. M., Mark, J. W.-K., & Caruso, A. 1971, *Ap Sp Sci*, 14, 52
- Lacey, C. G. 1984, *MNRAS*, 208, 687
- Lacey, C. G. 1991, in *Dynamics of Disc Galaxies*, ed. B. Sundelius (Gothenburg: Göteborgs University), 257
- Lacey, C. G., & Ostriker, J. P. 1985, *ApJ*, 299, 633
- Law, D. R., Johnston, K. V., & Majewski, S. R. 2005, *ApJ*, 619, 807
- Levine, E. S., Blitz, L., & Heiles, C. 2006, *ApJ*, 643, 881
- Li, C., Gadotti, D. A., Mao, S., & Kauffmann, G. 2009, *MNRAS*, 397, 726

- Lin, C. C., & Shu, F. H. 1966, *Proc Nat Acad Sci*, 55, 229
- Lin, D. N. C., & Tremaine, S. 1983, *ApJ*, 264, 364
- Lindblad, P. A. B., Lindblad, P. O., & Athanassoula, E. 1996, *A&A*, 313, 65
- Little, B., & Carlberg, R. G. 1991, *MNRAS*, 250, 161
- Louis, P. D., & Gerhard, O. E. 1988, *MNRAS*, 233, 337
- Lovelace, R. V. E. 1998, *A&A*, 338, 819
- Lovelace, R. V. E., & Hohlfield, R. G. 1978, *ApJ*, 221, 51
- Lovelace, R. V. E., Zhang, L., Kornreich, D. A. & Haynes, M. P. 1999, *ApJ*, 524, 634
- Lowe, S. A., Roberts, W. W., Yang, J., Bertin, G., & Lin, C. C. 1994, *ApJ*, 427, 184
- Lynden-Bell, D. 1965, *MNRAS*, 129, 299
- Lynden-Bell, D. 1979, *MNRAS*, 187, 101
- Lynden-Bell, D., & Kalnajs, A. J. 1972, *MNRAS*, 157, 1
- Maciejewski, W. 2006, *MNRAS*, 371, 451
- Maciejewski, W., & Sparke, L. S. 2000, *MNRAS*, 313, 745
- Maciejewski, W., Teuben, P. J., Sparke, L. S., & Stone, J. M. 2002, *MNRAS*, 329, 502
- Maoz, D., Barth, A. J., Ho, L. C., Sternberg, A., & Filippenko, A. V. 2001, *AJ*, 121, 3048
- Marinova, I., & Jogee, S. 2007, *ApJ*, 659, 1176
- Mark, J. W-K. 1974, *ApJ*, 193, 539
- Mark, J. W-K. 1976, *ApJ*, 203, 81
- Mark, J. W-K. 1977, *ApJ*, 212, 645
- Martinez-Valpuesta, I., & Shlosman, I. 2004, *ApJL*, 613, L29
- Mashchenko, S., Couchman, H. M. P., & Wadsley, J. 2006, *Nature*, 442, 539
- Masset, F., & Tagger, M. 1997, *A&A*, 322, 442
- Masters, K. L., Nichol, R. C., Hoyle, B., Lintott, C., Bamford, S., Edmondson, E. M., Fortson, L., Keel, W. C., Schawinski, K., Smith, A., & Thomas, D. 2010, [arXiv:1003.0449](https://arxiv.org/abs/1003.0449)
- Mathur, S. D. 1990, *MNRAS*, 243, 529
- Matthews, L. D. 2000, *AJ*, 120, 1764
- McMillan, P. J., & Dehnen, W. 2005, *MNRAS*, 363, 1205
- McMillan, P. J., & Dehnen, W. 2007, *MNRAS*, 378, 541
- Meidt, S. E., Rand, R. J., Merrifield, M. R., Debattista, V. P., & Shen, J. 2008, *ApJ*, 676, 899
- Meidt, S. E., Rand, R. J., & Merrifield, M. R. 2009, *ApJ*, 702, 277
- Méndez-Abreu, J., Sánchez-Janssen, R., & Aguerri, J. A. L. 2010, *ApJL*, 711, L61
- Merrifield, M. R., & Kuijken, K. 1999, *A&A*, 345, L47
- Merritt, D., & Sellwood, J. A. 1994, *ApJ*, 425, 551
- Merritt, D., & Stiavelli, M. 1990, *ApJ*, 358, 399-417
- Mestel, L. 1963, *MNRAS*, 126, 553
- Miller, R. H., Prendergast, K. H., & Quirk, W. J. 1970, *ApJ*, 161, 903
- Minchev, I., & Quillen, A. C. 2006, *MNRAS*, 368, 623
- Moster, B. P., Macciò, A. V., Somerville, R. S., Johansson, P. H., & Naab, T. 2010, *MNRAS*, 403, 1009
- Mould, J. 2005, *AJ*, 129, 698
- Mulder, W. A. 1983, *A&A*, 117, 9
- Nelson, R. W., & Tremaine, S. 1995, *MNRAS*, 275, 897
- Nieten, Ch., Neiningner, N., Guélin, M., Ungerechts, H., Lucas, R., Berkhuijsen, E. M., Beck, R., & Wielebinski, R. 2006, *A&A*, 453, 459
- Noguchi, M. 1987, *MNRAS*, 228, 635
- Nordström, B., Mayor, M., Andersen, J., Holmberg, J., Pont, F., Jørgensen, B. R., Olsen, E. H., Udry, S., & Mowlavi, N. 2004, *A&A*, 418, 989
- Norman, C. A., Sellwood, J. A., & Hasan, H. 1996, *ApJ*, 462, 114
- O'Neill, J. K., & Dubinski, J. 2003, *MNRAS*, 346, 251
- Oort, J. H. 1965, in *Stars and Stellar Systems, 5 Galactic Structure*, eds. A. Blaauw & M. Schmidt (Chicago: University of Chicago Press), 455
- Oort, J. H., Kerr, F. J., & Westerhout, G. 1958, *MNRAS*, 118, 379
- Ostriker, E. C., & Binney, J. J. 1989, *MNRAS*, 237, 785
- Ostriker, J. P., & Peebles, P. J. E. 1973, *ApJ*, 186, 467
- Papaloizou, J. C. B., & Lin, D. N. C. 1989, *ApJ*, 344, 645
- Patsis, P. A., Contopoulos, G., & Grosbol, P. 1991, *A&A*, 243, 373
- Patsis, P. A., Skokos, Ch., & Athanassoula, E. 2002, *MNRAS*, 337, 578
- Pérez, I. 2008, *A&A*, 478, 717
- Pérez, I., Fux, R., & Freeman, K. 2004, *A&A*, 424, 799
- Pfenniger, D., & Friedli, D. 1991, *A&A*, 252, 75
- Pfenniger, D., & Norman, C. 1990, *ApJ*, 363, 391
- Pichon, C., & Cannon, R. C. 1997, *MNRAS*, 291, 616
- Polyachenko, E. V. 2004, *MNRAS*, 348, 345
- Polyachenko, E. V. 2005, *MNRAS*, 357, 559
- Polyachenko, V. L. 1977, *Sov Ast Lett*, 3, 51
- Prendergast, K. H. 1962, in *Interstellar Matter in Galaxies*, ed. L. Woltjer (New York: Benjamin), 217
- Quillen, A. C. 2003, *AJ*, 125, 785
- Quillen, A. C., & Garnett, D. R. 2000, [arXiv:astro-ph/0004210](https://arxiv.org/abs/astro-ph/0004210)
- Quinn, T., & Binney, J. 1992, *MNRAS*, 255, 729
- Raboud, D., Grenon, M., Martinet, L., Fux, R., & Udry, S. 1998, *A&A*, 335, L61
- Rafikov, R. R. 2001, *MNRAS*, 323, 445
- Raha, N., Sellwood, J. A., James, R. A., & Kahn, F. D. 1991, *Nature*, 352, 411
- Rautiainen, P., & Salo, H. 1999, *A&A*, 348, 737
- Rautiainen, P., Salo, H., & Laurikainen, E. 2002, *MNRAS*, 337, 1233
- Rautiainen, P., Salo, H., & Laurikainen, E. 2008, *MNRAS*, 388, 1803
- Read, J. I., Lake, G., Agertz, O., & Debattista, V. P. 2008, *MNRAS*, 389, 1041

- Reese, A., Williams, T. B., Sellwood, J. A., Barnes, E. I., & Powell, B. A. 2007, *AJ*, 133, 2846
- Regan, M. W., Sheth, K., Teuben, P. J., & Vogel, S. N. 2002, *ApJ*, 574, 126
- Reid, I. N., Turner, E. L., Turnbull, M. C., Moun-tain, M., & Valenti, J. A. 2007, *ApJ*, 665, 767
- Reshetnikov, V., Battaner, E., Combes, F., & Jiménez-Vicente, J. 2002, *A&A*, 382, 513
- Reylé, C., Marshall, D. J., Robin, A. C., & Schultheisé, M. 2009, *A&A*, 495, 819
- Roberts, W. W., Huntley, J. M., & van Albada, G. D. 1979, *ApJ*, 233, 67
- Romano-Díaz, E., Shlosman, I., Hoffman, Y., & Heller, C. 2008a, *ApJL*, 685, L105
- Romano-Díaz, E., Shlosman, I., Heller, C., & Hoffman, Y. 2008b, *ApJL*, 687, L13
- Romeo, A. B. 1992, *MNRAS*, 256, 307
- Roškar, R., Debattista, V. P., Quinn, T. R., Stinson, G. S., & Wadsley, J. 2008, *ApJL*, 684, L79
- Roškar, R., Debattista, V. P., Stinson, G. S., Quinn, T. R., Kaufmann, T., & Wadsley, J. 2008, *ApJL*, 675, L65
- Rubin, V. C., Graham, J. A., & Kenney, J. D. P. 1992, *ApJL*, 394 L9
- Rybicki, G. B. 1972, in *IAU Colloq. 10, Gravitational N-body Problem*, ed. M. Lecar (Dordrecht: Reidel), 22
- Saha, K., Combes, J., & Jog, C. 2007, *MNRAS*, 382, 419
- Saha, K., de Jong, R., & Holwerda, B. 2009, *MNRAS*, 396, 409
- Salo, H., & Laurikainen, E. 1993, *ApJ*, 410, 586
- Sancisi, R. 1976, *A&A*, 53, 159
- Sancisi, R., Fraternali, F., Oosterloo, T., & van der Hulst, T. 2008, *A&A Rev.*, 15, 189
- Sandage, A., & Humphreys, R. M. 1980, *ApJL*, 236, L1
- Sanders, R. H., & Huntley, J. M. 1976, *ApJ*, 209, 53
- Sawamura, M. 1988, *PASJ*, 40, 279
- Schönrich, R., & Binney, J. 2009, *MNRAS*, 396, 203
- Schwarz, M. P. 1981, *ApJ*, 247, 77
- Seabroke, G. M., & Gilmore, G. 2007, *MNRAS*, 380, 1348
- Sellwood, J. A. 1980, *A&A*, 89, 296
- Sellwood, J. A. 1981, *A&A*, 99, 362
- Sellwood, J. A. 1985, *MNRAS*, 217, 127
- Sellwood, J. A. 1989a, *MNRAS*, 238, 115
- Sellwood, J. A. 1989b, in *Dynamics of Astrophysical Discs*, ed. J. A. Sellwood (Cambridge: Cambridge University Press), 155
- Sellwood, J. A. 1994, in *Galactic and Solar System Optical Astrometry*, ed. L. Morrison (Cambridge: Cambridge University Press), 156
- Sellwood, J. A. 1996a, in *IAU Symp. 169, Unsolved Problems of the Milky Way*, eds. L. Blitz & P. Teuben (Dordrecht: Kluwer), 31
- Sellwood, J. A. 1996b, *ApJ*, 473, 733
- Sellwood, J. A. 2000, in *Astrophysical Dynamics – in Commemoration of F. D. Kahn*, eds. D. Berry, D. Breitschwerdt, A. da Costa, & J. E. Dyson, *Ap Sp Sci*, 272, 31 (astro-ph/9909093)
- Sellwood, J. A. 2003, *ApJ*, 587, 638
- Sellwood, J. A. 2006, *ApJ*, 637, 567
- Sellwood, J. A. 2008a, *ApJ*, 679, 379
- Sellwood, J. A. 2008b, in *Formation and Evolution of Galaxy Disks*, eds. J. G. Funes SJ & E. M. Corsini (San Francisco: ASP 396), 341 (arXiv:0803.1574)
- Sellwood, J. A. 2010, *MNRAS*, submitted (arXiv:1001.5197)
- Sellwood, J. A., & Binney, J. J. 2002, *MNRAS*, 336, 785
- Sellwood, J. A., & Carlberg, R. G. 1984, *ApJ*, 282, 61
- Sellwood, J. A., & Debattista, V. P. 2006, *ApJ*, 639, 868
- Sellwood, J. A., & Debattista, V. P. 2009, *MNRAS*, 398, 1279
- Sellwood, J. A., & Evans, N. W. 2001, *ApJ*, 546, 176
- Sellwood, J. A., & Kahn, F. D. 1991, *MNRAS*, 250, 278
- Sellwood, J. A., & Lin, D. N. C. 1989, *MNRAS*, 240, 991
- Sellwood, J. A., & Merritt, D. 1994, *ApJ*, 425, 530
- Sellwood, J. A., & Moore, E. M. 1999, *ApJ*, 510, 125
- Sellwood, J. A., Nelson, R. D., & Tremaine, S. 1998, *ApJ*, 506, 590
- Sellwood, J. A., & Sparke, L. S. 1988, *MNRAS*, 231, 25P
- Sellwood, J. A., & Valluri, M. 1997, *MNRAS*, 287, 124
- Sellwood, J. A., & Wilkinson, A. 1993, *Rep Prog Phys*, 56, 173
- Shakura, N. I., & Sunyaev, R. A. 1973, *A&A*, 24, 337
- Shen, J., & Debattista, V. P. 2009, *ApJ*, 690, 758
- Shen, J., & Sellwood, J. A. 2004, *ApJ*, 604, 614
- Shen, J., & Sellwood, J. A. 2006, *MNRAS*, 370, 2
- Sheth, K. et al. 2008, *ApJ*, 675, 1141
- Shetty, R., Vogel, S. N., Ostriker, E. C., & Teuben, P. J. 2007, *ApJ*, 665, 1138
- Shiidsuke, K., & Ida, S. 1999, *MNRAS*, 307, 737
- Shlosman, I., Frank, J., & Begelman, M. C. 1989, *Nature*, 338, 45
- Shu, F. H., Tremaine, S., Adams, F. C., & Ruden, S. P. 1990, *ApJ*, 358, 495
- Skokos, Ch., Patsis, P. A., & Athanassoula, E. 2002, *MNRAS*, 333, 847
- Soderblom, D. R. 2010, *ARAA*, to appear (arXiv:1003.6074)
- Sparke, L. S., & Casertano, S. 1988, *MNRAS*, 234, 873
- Sparke, L. S., & Sellwood, J. A. 1987, *MNRAS*, 225, 653
- Spitzer, L. 1942, *ApJ*, 95, 329
- Spitzer, L., & Schwarzschild, M. 1953, *ApJ*, 118, 106
- Stanghellini, L., & Haywood, M. 2010, *ApJ*, 714, 1096
- Stewart, K. R., Bullock, J. S., Wechsler, R. H., Maller, A. H., & Zentner, A. R. 2008, *ApJ*, 683, 597
- Syget, J. F., Tagger, M., Athanassoula, E., & Pellat, R. 1988, *MNRAS*, 232, 733

- Tagger, M., Sygnet, J. F., Athanassoula, E., & Pellat, R. 1987, *ApJL*, 318, L43
- Thomasson, M., Elmegreen, B. G., Donner, K. J., & Sundelius, B. 1990, *ApJL*, 356, L9
- Toomre, A. 1964, *ApJ*, 139, 1217
- Toomre, A. 1966, in *Geophysical Fluid Dynamics, notes on the 1966 Summer Study Program at the Woods Hole Oceanographic Institution*, ref. no. 66-46
- Toomre, A. 1969, *ApJ*, 158, 899
- Toomre, A. 1981, in *The Structure and Evolution of Normal Galaxies*, eds. S. M. Fall, & D. Lynden-Bell (Cambridge: Cambridge University Press), 111
- Toomre, A. 1983, in *IAU Symposium 100, Internal Kinematics and Dynamics of Galaxies*, ed. E. Athanassoula (Dordrecht: Reidel), 177
- Toomre, A. 1989, in *Dynamics of Astrophysical Discs*, ed. J. A. Sellwood (Cambridge: Cambridge University Press), 153
- Toomre, A. 1990, in *Dynamics & Interactions of Galaxies*, ed. R. Wielen (Berlin, Heidelberg: Springer), 292
- Toomre, A. 1995, unpublished notes
- Toomre, A., & Kalnajs, A. J. 1991, in *Dynamics of Disc Galaxies*, ed. B. Sundelius (Gothenburg: Göteborgs University), 341
- Tóth, G., & Ostriker, J. P. 1992, *ApJ*, 389, 5
- Tremaine, S. 2005, *ApJ*, 625, 143
- Tremaine, S., & Weinberg, M. D. 1984a, *ApJL*, 282, L5
- Tremaine, S., & Weinberg, M. D. 1984b, *MNRAS*, 209, 729
- Tsoutsis, P., Kalapotharakos, C., Efthymiopoulos, C., & Contopoulos, G. 2009, *A&A*, 495, 743
- Valenzuela, O., & Klypin, A. 2003, *MNRAS*, 345, 406
- Vandervoort, P. O. 1970, *ApJ*, 161, 87
- Vauterin, P., & Dejonghe, H. 1996, *A&A*, 313, 465
- Velázquez, H., & White, S. D. M. 1999, *MNRAS*, 304, 254
- Villa-Vargas, J., Shlosman, I., & Heller, C. 2009, *ApJ*, 707, 218
- Villumsen, J. V. 1985, *ApJ*, 290, 75
- Voglis, N., Harsoula, M., & Contopoulos, G. 2007, *MNRAS*, 381, 757
- Wada, K. 2001, *ApJL*, 559, L41
- Walker, I. R., Mihos, J. C., & Hernquist, L. 1996, *ApJ*, 460, 121
- Weinberg, M. D. 1985, *MNRAS*, 213, 451
- Weinberg, M. D. 1991, *ApJ*, 373, 391
- Weinberg, M. D. 1994, *ApJ*, 421, 481
- Weinberg, M. D., & Blitz, L. 2006, *ApJL*, 641, L33
- Weinberg, M. D., & Katz, N. 2002, *ApJ*, 580, 627
- Weinberg, M. D., & Katz, N. 2007, *MNRAS*, 375, 425
- Weiner, B. J. 2004, in *IAU Symp. 220, Dark Matter in Galaxies*, eds. S. Ryder, D. J. Pisano, M. Walker, & K. C. Freeman (Dordrecht: Reidel), 35
- Weiner, B. J., Sellwood, J. A., & Williams, T. B. 2001, *ApJ*, 546, 931
- Wielen, R. 1977, *A&A*, 60, 263
- Yoachim, P., & Dalcanton, J. J. 2006, *AJ*, 131, 226
- Zang, T. A. 1976, PhD thesis, MIT
- Zang, T. A., & Hohl, F. 1978, *ApJ*, 226, 521
- Zánmar Sánchez, R., Sellwood, J. A., Weiner B. J., & Williams, T. B. 2008, *ApJ*, 674, 797
- Zhang, X. 1996, *ApJ*, 457, 125
- Zhang, X. 1998, *ApJ*, 499, 93
- Zibetti, S., Charlot, S., & Rix, H.-W. 2009, *MNRAS*, 400, 1181

19 Mass Distribution and Rotation Curve in the Galaxy

Yoshiaki Sofue

Institute of Astronomy, The University of Tokyo, Tokyo, Japan

Department of Physics, Meisei University, Tokyo, Japan

1	<i>Introduction</i>	987
2	<i>Rotation Curves</i>	990
2.1	Measurements of Galactic Rotation	990
2.1.1	Rotation Velocities from Distance, Radial Velocity, and Proper Motion	990
2.1.2	Terminal-Velocity Method: Rotation Curve Inside Solar Circle	993
2.1.3	Ring Thickness Method	993
2.2	Rotation Curve of the Galaxy	993
2.2.1	Rotation in the Galactic Center	994
2.3	Measurements of Rotation Velocities in External Galaxies	996
2.3.1	H α and Optical Measurements	996
2.3.2	Radio Lines: HI, CO, and Masers	997
2.3.3	Centroid Velocity Methods	997
2.3.4	Terminal-Velocity Method	999
2.3.5	Iteration Method	999
2.3.6	Three-Dimensional Cube Method for the Future	1000
2.4	Rotation Curves of Spiral Galaxies	1000
2.4.1	Sa, Sb, Sc Galaxies	1002
2.4.2	Barred Galaxies	1003
2.4.3	Dwarf Galaxies	1003
2.4.4	Irregular Galaxies	1003
3	<i>Galactic Mass Distribution</i>	1004
3.1	Approximate Mass Distribution	1004
3.1.1	Flat Rotation and Isothermal Mass Distribution	1004
3.1.2	Vertical Mass Distribution Near the Sun	1004
3.2	Decomposition Methods	1006
3.3	Decomposition into Bulge, Disk, and Halo	1006
3.3.1	de Vaucouleurs Bulge	1007
3.3.2	M/L Ratio in Bulge	1009
3.3.3	Exponential Disk	1009
3.3.4	Dark Halo	1010
3.4	Galactic Mass Parameters	1010
3.5	Miyamoto–Nagai Potential	1011
3.6	Direct Method	1013

3.6.1	Spherical Mass Distribution	1013
3.6.2	Flat-Disk Mass Distribution	1015
3.6.3	Verification Using the Miyamoto–Nagai Potential	1015
3.7	Direct Mass Distribution in the Galaxy	1017
3.8	Direct Mass Distributions in Spiral Galaxies	1018
3.9	Distribution of Interstellar Gas in the Galaxy	1019
4	<i>Dark Halo</i>	1021
4.1	Dark Halo in the Milky Way	1021
4.1.1	Isothermal Halo Model	1021
4.1.2	NFW and Burkert Profiles	1022
4.1.3	Dark Halo Contribution in the Inner Galaxy	1024
4.2	Pseudo Rotation Curve of the Local Group	1025
4.3	Mass of the Galaxy Embedded in the Dark Halo	1026
4.4	The Galaxy, M31, and Local Group	1027
4.5	Dark Halos in Galaxies	1029
4.6	Mass-to-Luminosity Ratio	1029
5	<i>Smaller Mass Structures</i>	1030
5.1	Spiral Arms and Rotation Dips	1030
5.2	Bar	1032
5.3	Massive Central Component and Black Hole	1033
6	<i>Summary</i>	1034
	<i>References</i>	1035

Abstract: The mass distribution in the Galaxy is determined by dynamical and photometric methods. The dynamical method is based on the Virial theorem, and calculates the mass from kinematical data such as rotation velocities, velocity dispersions, and motions of satellite galaxies. Rotation curves are the major tool for determining the dynamical mass distribution in the Milky Way and spiral galaxies. The photometric (statistical) method utilizes luminosity profiles from optical and infrared observations, and assumes empirical values of the mass-to-luminosity (M/L) ratio to convert the luminosity to mass. This method is convenient to separate the mass components such as bulge and disk, while the uncertainty is large due to ambiguous M/L ratio arising from the variety of stellar populations. Also, the methods cannot detect the dark matter that dominates in the outer regions and central black holes.


In this chapter the dynamical method is described in detail, and rotation curves and mass distribution in the Milky Way and nearby spiral galaxies are presented. The dynamical method is further categorized into two methods: the decomposition method and direct method. The former fits the rotation curve by calculated curve assuming several mass components such as a bulge, disk, and halo, and adjusts the dynamical parameters of each component. Explanations are given of the mass profiles as the de Vaucouleurs law, exponential disk, and dark halo profiles inferred from numerical simulations. Another method is the direct method, with which the mass distribution can be directly calculated from the data of rotation velocities without employing any mass models. Some results from both methods are presented, and the Galactic structure is discussed in terms of the mass. Rotation curves and mass distributions in external galaxies are also discussed, and the fundamental mass structures are shown to be universal.

Keywords: Bulge, Dark halo, Dark matter, Dynamics, Galactic disk, Galactic halo, Galaxies, Local group, Mass distribution, Rotation curve, The galaxy

1 Introduction

The mass of a galaxy and its distribution are obtained by two ways: one is to use photometric data such as luminosity profiles assuming the mass-to-luminosity (M/L) ratio and the other is to apply dynamics to kinematical data based on the Virial theorem. The photometric method has been used to estimate the mass of the Galaxy early in the beginning of the twentieth century by counting stellar density in the solar neighborhood. The density was multiplied by the total volume of the Galaxy to calculate the number of solar-mass stars. This estimation already gave an approximate mass on the order of $10^{11} M_{\odot}$ of the Galaxy (Fich and Tremaine 1991).

Most of the luminous mass is occupied by stars, and the rest $\sim 10\%$ is by interstellar gases. Hence, the stellar luminosity distribution roughly represents the luminous mass distribution. The distribution of stellar mass can be obtained from optical and infrared surface photometry of spiral galaxies, by assuming the mass-to-luminosity (M/L) ratio. Particularly, near-infrared K band ($2.2 \mu\text{m}$) images are useful to obtain an approximate distribution of stars occupying most of the luminous mass. Infrared photometry along the Galactic plane has been used to derive the mass distribution in the Galaxy. Although the photometric method is convenient to approximately map the luminous mass, it varies with the employed M/L ratio. In order to determine the M/L ratio, measurement of mass independent of luminosity observation is in any way necessary beforehand. Besides the ambiguity of the assumed M/L ratio, it cannot give information about the dark matter.

Nevertheless, the luminous matter distribution is helpful to overview the mass components in the Galaxy.  *Figure 19-1* shows a schematic view of a spiral galaxy, which is composed of the bulge, disk, and halo. The luminous halo is dominated by globular clusters and high-temperature gas, whereas the mass is dominated by dark matter. In addition to these luminous components, a galaxy nests a central massive object, or a black hole, and a massive core in the bulge. The entire galaxy is surrounded by a huge dark halo composed of dark matter.

For the mass including the dark matter and invisible masses like black holes, it is necessary to apply the dynamical method using kinematical data. The dynamical mass is calculated based on the Virial theorem that the kinetic and gravitational energies are in equilibrium for a dynamically relaxed system. For a rotating object such as the galactic disk, the balance between the gravitational and centrifugal forces is applied.

The dynamical mass of the Galaxy on the order of $\sim 10^{11} M_{\odot}$ inside the solar circle was already calculated early in the 1950s (Oort 1958) when the circular orbit of the local standard of rest (LSR) was obtained in terms of the galactic-centric distance R_0 and rotation velocity V_0 (see Fich and Tremaine 1991 for a review). For a standard set of the parameters of $R_0 = 8$ kpc and $V_0 = 200$ km s $^{-1}$, they yield the most fundamental quantity, the mass inside the solar circle on an assumption of spherical distribution, to be on the order of

$$M_0 = R_0 V_0^2 / G = 7.44 \times 10^{10} M_{\odot} \sim 10^{11} M_{\odot}, \quad (19.1)$$

where G is the gravitational constant. Although this approximate estimation is not far from the true value, the mass distribution in the Galaxy, as well as those in any galaxies, is not simply spherical, and it is principally derived by analyzing the rotation curves on the assumption that the centrifugal force of the circular motion is balancing with the gravitational force in a spheroid or a disk. Hence, the first step to derive the mass and mass distributions in galaxies is to obtain the rotation curves. Given a rotation curve, it is deconvolved to several components representing the mass distribution using various methods.

The inner rotation curve is simply measured by the terminal (tangential) – velocity method applied to radio line observations such as the HI and CO lines. The mass profiles of the galactic disk and bulge were thus known since the 1970s. The central mass condensation and the nuclear massive black hole have been measured since the 1980s when kinematics of interstellar gas and stars close to Sgr A * , our Galaxy's nucleus, were measured by infrared observations. For the total mass of the Galaxy including the outermost regions, they had to wait until an outer rotation curve and detailed analyses of motions of member galaxies in the Local Group were obtained. It was only recent when the total mass was estimated to be $\sim 3 \times 10^{11} M_{\odot}$ including the dark halo up to ~ 150 kpc by considering the outer rotation curve and motions of satellite galaxies.

The galactic constants, which are the galactocentric distance of the Sun R_0 and the rotation velocity V_0 of the local standard of rest (LSR) around the Galactic Center (GC), namely the three-dimensional position and motion of the LSR, are the most fundamental parameters to derive the mass and its distribution in the Galaxy. The LSR is defined as the coordinates with its origin at the Sun and rotating on a circular orbit around the GC after correcting for the solar motion. The galactocentric distance R_0 has been determined by various methods, which lies in the range of 7 and 9 kpc (Reid 1993). Given the value of R_0 , the rotation velocity V_0 is determined by using the Oort's constants A and B as

$$V_0 = (A - B)R_0, \quad (19.2)$$

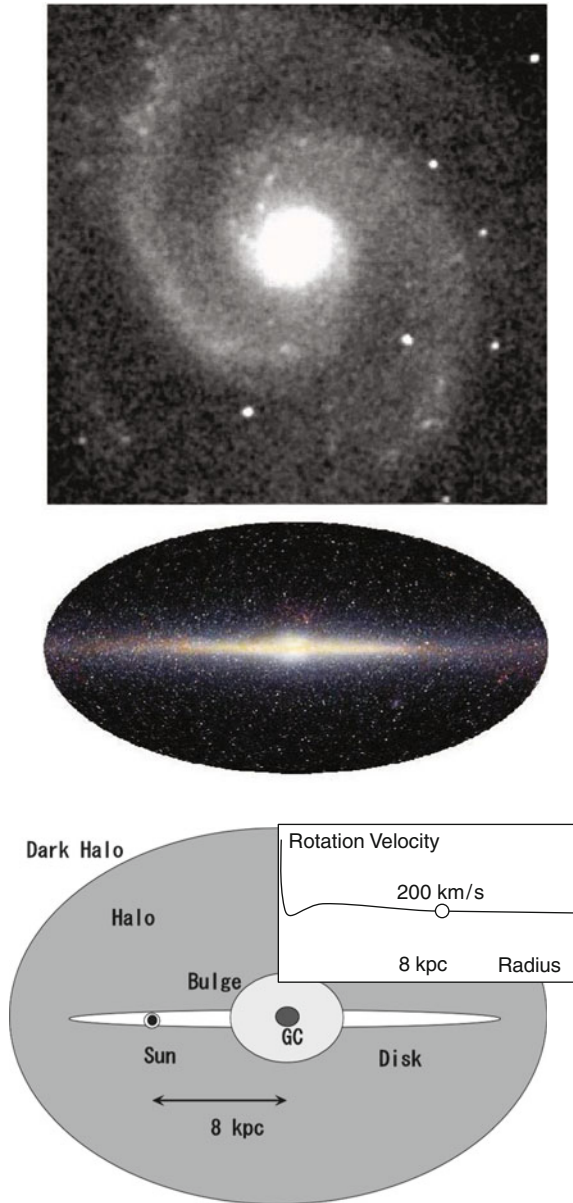


Fig. 19-1

(Top) Near-infrared (K band, $2.2 \mu\text{m}$) image of a face-on spiral galaxy M51 from 2MASS survey (The Milky Way is rotating clockwise as seen from the North Galactic Pole, showing trailing arms in opposite sense to this image.) (Middle) Whole sky near-infrared view of the Milky Way from COBE, showing the Galaxy's edge-on view. These images approximately represent the distributions of stars, therefore, the luminous mass. They are embedded in a massive dark halo of comparable mass. (Bottom) Schematic galactic structure. The Galaxy consists of the nucleus nesting a massive black hole, high-density massive core, bulge, disk, halo objects, and a dark matter halo. Inserted is a schematic rotation curve

where

$$A = \frac{1}{2} \left(\frac{V}{R} - \frac{dV}{dR} \right)_{R_0} \quad (19.3)$$

and

$$B = -\frac{1}{2} \left(\frac{V}{R} - \frac{dV}{dR} \right)_{R_0}, \quad (19.4)$$

with R and V being galactocentric distance and rotation velocity of stars in the solar neighborhood. Here the values $R_0 = 8$ kpc and $V_0 = 200$ km s⁻¹ are adopted, which yield approximate mass of the Galaxy inside the solar circle of $M_0 = 7.44 \times 10^{10} M_\odot$ as in (► 19.1).

In this chapter the dynamical methods are described and are applied to determination of the mass distribution in the Galaxy using the rotation curve. The description will be made of two parts: one for rotation curve, and another for deconvolution into mass components and density profiles. Mass distributions in external galaxies and their rotation curves are also touched upon. More detailed description of individual methods and analyses may be referenced to various papers in the cited literature. ► Section 2 is partly based on the review by Sofue and Rubin (2001). ► Sections 3 and ► 4 are based on the studies of galactic mass models by Sofue et al. (2009) and Sofue (2009).

2 Rotation Curves

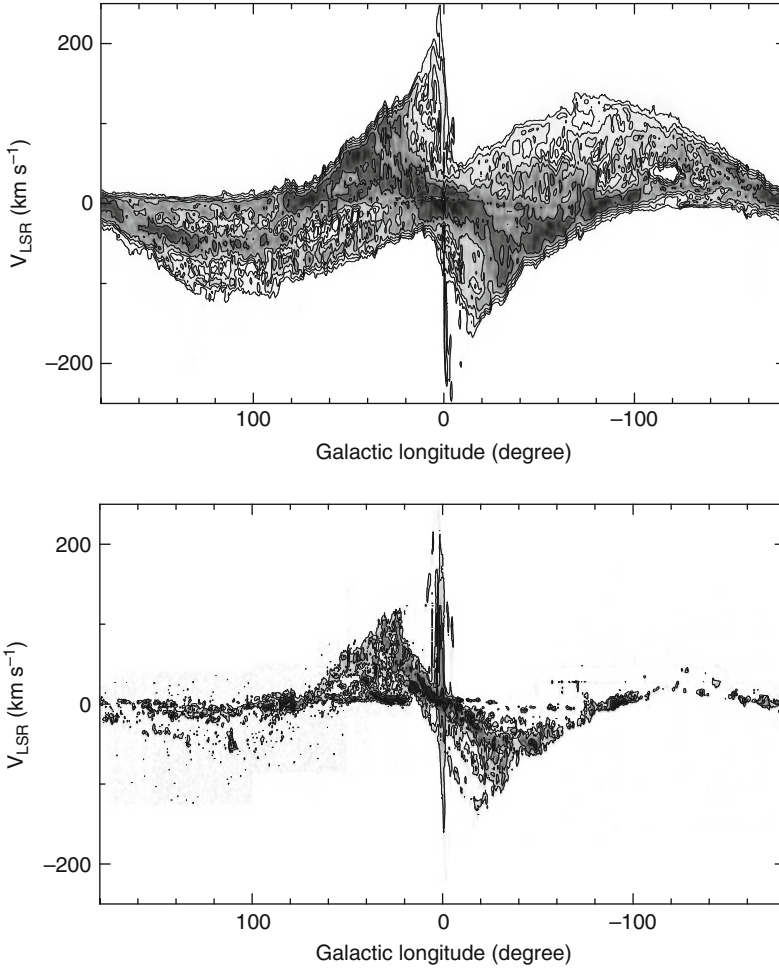
2.1 Measurements of Galactic Rotation

The rotation of the Milky Way is clearly seen in longitude-radial velocity (LV) diagrams along the galactic plane, where spectral line intensities are plotted on the (l, V_{lsr}) plane as shown in ► Fig. 19-2, which shows the observed LV diagrams in the $\lambda 21$ -cm HI and $\lambda 2.6$ -mm CO emission lines.

The positive- and negative-velocity envelopes (terminal velocities) at $0 < l < 90^\circ$ and $270 < l < 360^\circ$, respectively, are used for determining the inner rotation curve inside the solar circle at $R \leq R_0$. This terminal (tangential)-velocity method is applied to HI- and CO-line observations for the inner Galaxy (Burton and Gordon 1978; Clemens 1985; Fich et al. 1989). In order to derive outer rotation curve beyond the solar circle, optical distances and velocities of OB stars are combined with CO-line velocities (Blitz et al. 1982; Demers and Battinelli 2007). HI thickness method is useful to obtain rotation curve of the entire disk (Merrifield 1992; Honma and Sofue 1997). High-accuracy measurements of parallax and proper motions of maser sources and Mira variable stars using VLBI technique are providing an advanced tool to derive a more accurate rotation curve (Honma et al. 2007).

2.1.1 Rotation Velocities from Distance, Radial Velocity, and Proper Motion

Given the galactic constants R_0 and V_0 , rotation velocity $V(R)$ in the galactic disk can be obtained as a function of galactocentric distance R by measuring the distance r radial velocity v_r and/or perpendicular velocity $v_p = \mu r$ of an object, where μ is the proper motion (► Fig. 19-3). The velocity vector of a star at any position in the Galaxy is determined by observing its three-dimensional position (r, l, b) and its motion (v_r, v_p) , where v_r is the radial velocity and v_p the perpendicular velocity with $v_p = \mu r$ with μ being the proper motion on the sky.



■ Fig. 19-2

Longitude-radial velocity ($l-V_{\text{LSR}}$) diagram of the $\lambda 21\text{-cm}$ HI-line emission (top: Nakanishi 2007) and $\lambda 2.6\text{-mm}$ CO (bottom: Dame et al. 1987) lines along the galactic plane

The galactocentric distance R is calculated from the position of the object (l, b, r) and R_0 as

$$R = (r^2 + R_0^2 - 2rR_0\cos l)^{1/2}. \quad (19.5)$$

Here, the distance r to the object must be measured directly by trigonometric (parallax) method, or indirectly by spectroscopic measurements.

If the orbit of the star is assumed to be circular in the galactic plane, the rotation velocity $V(R)$ may be obtained by measuring one of the radial velocity or proper motion. The rotation velocity $V(R)$ is related to the radial velocity v_r as

$$V(R) = \frac{R}{R_0} \left(\frac{v_r}{\sin l} + V_0 \right). \quad (19.6)$$

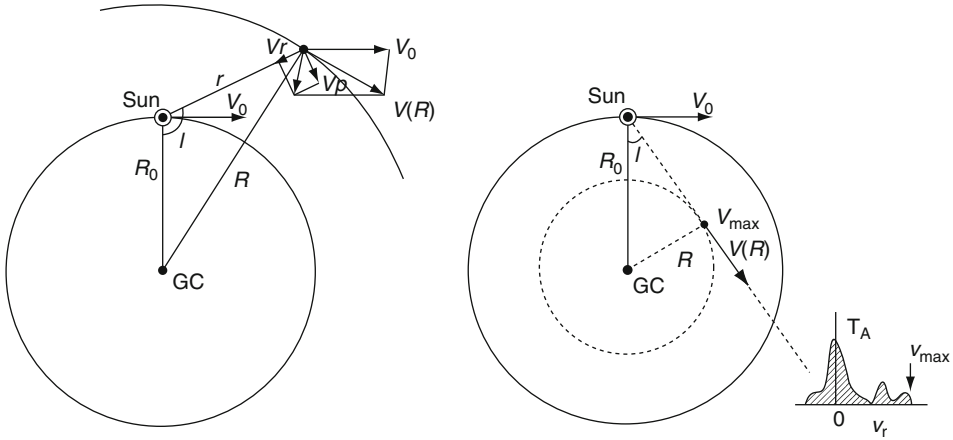


Fig. 19-3

(Left) Rotation velocity at any point in the galactic plane is obtained by measuring the distance r and either radial velocity v_r or perpendicular velocity $v_p = \mu r$, where μ is the proper motion. (Right) Rotation curve inside the solar circle ($R < R_0$; dashed circle) is obtained by measuring the terminal radial velocity v_{max} at the tangent point, where the GC distance is given by $R = R_0 \sin l$

Alternatively, the rotation velocity is determined by measuring the proper motion μ as

$$V(R) = -\frac{R}{s}(v_p + V_0 \cos l), \tag{19.7}$$

where

$$s = r - R_0 \cos l. \tag{19.8}$$

The method using radial velocity has been traditionally applied to various stellar objects. Star-forming regions are most frequently used to determine the rotation curve beyond the solar circle. In this method, distances r of OB stars are measured from their distance modulus from the apparent magnitude after correction for extinction and absolute luminosity by the star's color and spectral type. Then, the star's distance r is assumed to be the same as that of its associated molecular cloud and/or HII region whose radial velocity is obtained by observing the Doppler velocity of molecular lines and/or recombination lines. In this method, the error in the distance is large, which results in the large scatter in the obtained outer rotation curve, as seen in Fig. 19-5.

Accurate measurements of rotation velocities have been obtained, combining the spectroscopic and VLBI techniques by observing both the proper motion and radial velocity. Trigonometric (parallax) measurements of maser-line radio sources are used to determine the distance r and the proper motion $v_p (= r\mu)$, and the radial velocity of the same source v_r is measured by radio spectroscopic observations of the maser line. By this method, an accurate velocity has been obtained on the outer rotation curve as plotted in Fig. 19-5 by a big dot (Honma et al. 2007).

2.1.2 Terminal-Velocity Method: Rotation Curve Inside Solar Circle

From spectral profiles of interstellar gases as observed in the HI 21-cm and/or CO 2.6-mm emission lines in the first quadrant of the galactic plane ($0 < l < 90^\circ$), it is known that the gases within the solar circle show positive radial velocities, and those outside the solar circle have negative velocities because of the motion of the Sun (► *Figs. 19-2* and ► *19-3*). Maximum positive velocity is observed at the tangent point, at which the line of sight is tangential to the radius as indicated by the dashed lines in ► *Fig. 19-3*. This maximum radial velocity $v_{r \max}$ is called the terminal velocity or the tangent-point velocity. Using this terminal velocity, the rotation velocity $V(R)$ is simply calculated by

$$V(R) = v_{r \max} + V_0 \sin l, \quad (19.9)$$

and the galactocentric distance is given by

$$R = R_0 \sin l. \quad (19.10)$$

The rotation curve $V(R)$ in the first and fourth quadrants of the disk is thus determined by observing terminal velocities at various longitudes at $0 < l < 90^\circ$ and $270 < l < 360^\circ$, and therefore, at various R . In the fourth quadrant at $270 < l < 360^\circ$ the terminal velocities have negative values.

2.1.3 Ring Thickness Method

The HI-disk thickness method utilizes apparent width of an annulus ring of HI disk in the whole Galaxy (Merrifield 1992; Honma and Sofue 1997). This method yields annulus-averaged rotation velocity in the entire galactic disk. The method is illustrated in ► *Fig. 19-4*: The apparent latitudinal angle Δb of the HI disk along an annulus ring of radius R varies with longitude as

$$\Delta b = \arctan \left(\frac{z_0}{R_0 \cos l + \sqrt{R^2 - R_0^2 \sin^2 l}} \right). \quad (19.11)$$

with

$$v_r = W(R) \sin l, \quad (19.12)$$

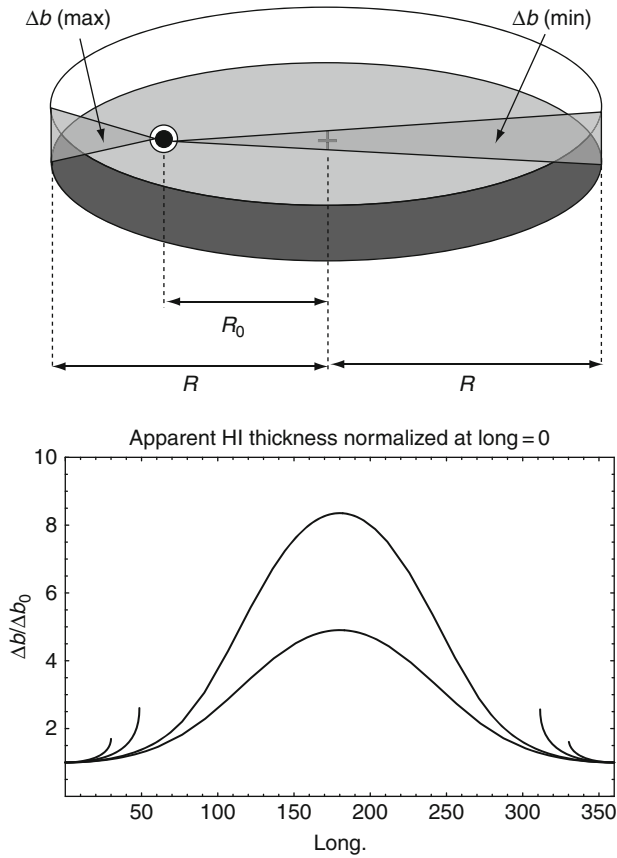
where

$$W(R) = \left[V(R) \frac{R_0}{R} - V_0 \right]. \quad (19.13)$$

The amplitude of Δb normalized by its value at $l = 180^\circ$ plotted against longitude l is uniquely related to the galactocentric distance R , which is as a function of $V(R)$ and is related to v_r as above equations. This method utilizes the entire HI disk, so that they obtained rotation curve manifests an averaged kinematics of the Galaxy. Therefore, it is informative for more global rotation curve compared to the measurements of individual stars or the tangent-point method.

2.2 Rotation Curve of the Galaxy

The entire rotation curve of the Galaxy is obtained by combining the observed rotation velocities from the various methods. However, when they calculate the rotation velocities from the



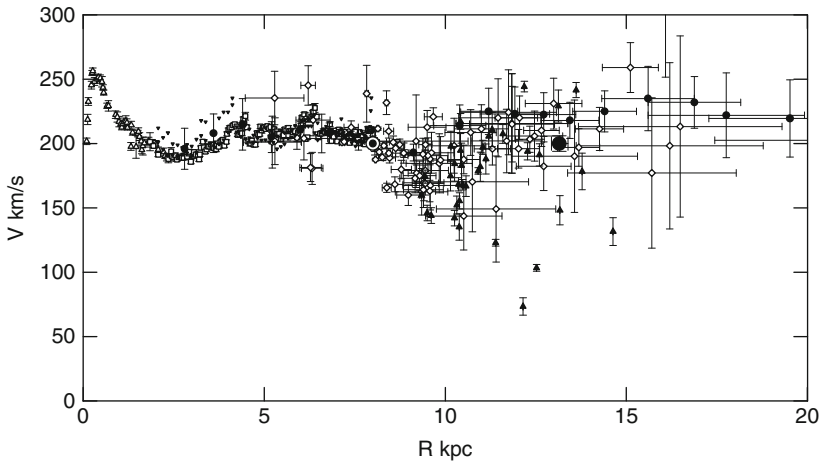
■ Fig. 19-4

Apparent thickness Δb of an annulus ring at R with the rotation velocity $V(R)$ varies with the longitude with the shape and relative amplitude being depending on the ring radius R . Indicated lines correspond to $R = 4, 6, 10,$ and 12 kpc

data, the observers often adopt different parameters. This had led to rotation curves of the Galaxy in different scaling both in R and $V(R)$. In order to avoid this inconvenience, a unified rotation curve was obtained by integrating the existing data by recalculating the distances and velocities for a nominal set of the galactocentric distance and the circular velocity of the Sun as $(R_0, V_0) = (8.0 \text{ kpc}, 200 \text{ km s}^{-1})$. ● *Figure 19-5* shows a plot of compiled data points for the rotation velocities (Sofue et al. 2009), and ● *Fig. 19-6* is a rotation curve fitted to the observed points, as described in the next subsections. Recent high-accuracy measurement by Honma et al. (2007) using VERA, an exact data point was given on the outer rotation curve at $(R, V(R)) = (13.15 \pm 0.22 \text{ kpc}, 200 \pm 6 \text{ km s}^{-1})$.

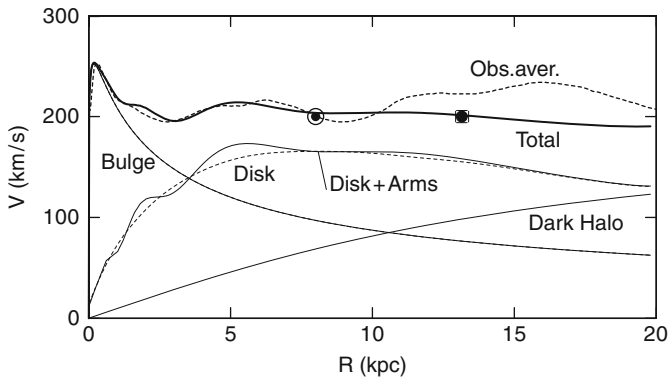
2.2.1 Rotation in the Galactic Center

The Galaxy provides a unique opportunity to derive a high-resolution central rotation curve (Gilmore et al. 1990). Proper-motion studies in the near infrared have revealed individual



■ Fig. 19-5

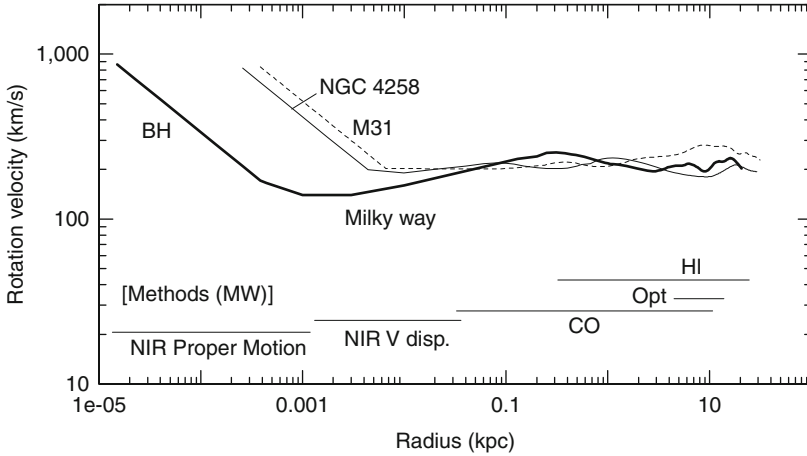
Observed circular velocities representing the rotation curve of the Galaxy (Sofue et al. 2009). Plot was obtained using HI tangent-velocity data (Burton and Gordon 1978; Fich et al. 1989); CO tangent-velocity data (Clemens 1985); CO cloud + HII region + OB stars (Blitz et al. 1982; Fich et al. 1989); Late type stars (Demers and Battinelli 2007); HI thickness method (Honma and Sofue 1997). Big circle at 13.1 kpc is from VERA by parallax, radial, and proper motions (Honma et al. 2007). All data have been converted to $(R_0, V_0) = (8.0, 200.0 \text{ km s}^{-1})$



■ Fig. 19-6

Calculated rotation curve including the bulge, disk, spiral arms, and dark halo. The big dot denotes the observed result from VERA (Honma et al. 2007). The pure disk component is also indicated by the *thin dashed line*. The *thick dashed line* indicates a simply averaged observed rotation curve taken from Sofue et al. (1999)

orbits of stars within the central 0.1 pc, and the velocity dispersion increases toward the center, indicating the existence of a massive black hole of mass $3 \times 10^6 M_{\odot}$ (Genzel et al. 1997, 2000; Ghez et al. 1998). Outside the very nuclear region, radial velocities of OH and SiO maser lines from IR stars in the Galactic Center region are used to derive the velocity dispersion and mean rotation (Lindqvist et al. 1992). SiO masers from IRAS sources in the central bulge have been



■ Fig. 19-7

Logarithmic rotation curve of the Galaxy (Sofue and Rubin 2001). For comparison those for two spiral galaxies are indicated. Innermost rotation velocities are Keplerian velocities calculated for the massive black holes. Observational methods for the Milky Way are shown by *horizontal lines*

used to study the kinematics, and the mean rotation of the bulge was found to be in solid body rotation of the order of 100 km s^{-1} (Deguchi et al. 2000; Izumiura et al. 1999). The very central rotation curve controlled by the black hole is Keplerian. In order to overview the entire rotation characteristics from the nucleus to the outer edge, the rotation curve is shown in a logarithmic scale in [Fig. 19-7](#).

2.3 Measurements of Rotation Velocities in External Galaxies

Rotation curves are also the major tool to derive the mass distribution in external galaxies, and are usually derived from Doppler velocities of the emission lines such as $\text{H}\alpha$, [NII], HI, and CO emission lines, which are particularly useful to calculate the mass distribution because of the small velocity dispersions of $\sim 10 \text{ km s}^{-1}$ compared to rotation velocities of $\sim 200 \text{ km s}^{-1}$. This allows us to neglect the pressure term in the equation of motion for calculating the mass distribution. Reviews on rotation curves of disk galaxies have been given by various authors (Ashman 1992; Persic and Salucci 1997; Persic et al. 1996; Sofue and Rubin 2001; Trimble 1987). On the other hand, spheroidal galaxies have higher velocity dispersions that are measured from stellar absorption lines (Binney 1982; de Zeeuw and Franx 1991; Faber and Gallagher 1979).



2.3.1 $\text{H}\alpha$ and Optical Measurements

In optical observations, long slit spectra are most often used to deduce the rotation curve of a galaxy from emission lines (Rubin et al. 1982, 1985; Mathewson et al. 1992; Mathewson and Ford 1996; Amram et al. 1995). The $\text{H}\alpha$, [NII], and [SII] emission lines are most commonly used. For a limited number of nearby galaxies, rotation curves can be produced from velocities

of individual HII regions in galactic disks (Rubin and Ford 1970, 1983; Zaritsky et al. 1997). In early galaxies, planetary nebulae are valuable tracers of the velocity fields in outer disks, where the optical light is faint. Fabry–Perot spectrometers are used to derive the $H\alpha$ velocity fields. Velocity fields cover the entire disk and yield more accurate rotation curve than the slit observations, while it requires more sophisticated analyses.

2.3.2 Radio Lines: HI, CO, and Masers

Radio wave emission lines are useful for studying the whole galactic disk, because the extinction is negligible even in the central regions, where the optical lines are often subject to strong absorption by the interstellar dusts. The λ 21-cm HI line is a powerful tool to obtain kinematics of the outer rotation curves, in part because its radial extent is greater than that of the visible disk (Bosma 1981a, b; van der Kruit and Allen 1978), while it is often weak or absent in the central regions.

The CO rotational transition lines in the millimeter wave range at 2.63 and 1.32 mm are valuable in studying rotation kinematics of the inner disks not only for the concentration of the molecular gas toward the center (Sofue 1996, 1997). Edge-on and high-inclination galaxies are particularly useful for rotation curve analysis in order to minimize the uncertainty arising from inclination corrections, for which extinction-free radio measurements are crucial. Another advantage of CO spectroscopy is its high spatial and velocity resolutions by interferometric observations (Sargent and Welch 1993; Sofue and Rubin 2001).  *Figure 19-8* illustrates by a simple simulation that the inner rotation curves is better traced by CO lines, while the outer rotation curve is well obtained from HI.  *Figure 19-9* shows a position-velocity (PV) diagram obtained for the edge-on galaxy NGC 3079 in the HI- and CO-line interferometer observations.

2.3.3 Centroid Velocity Methods

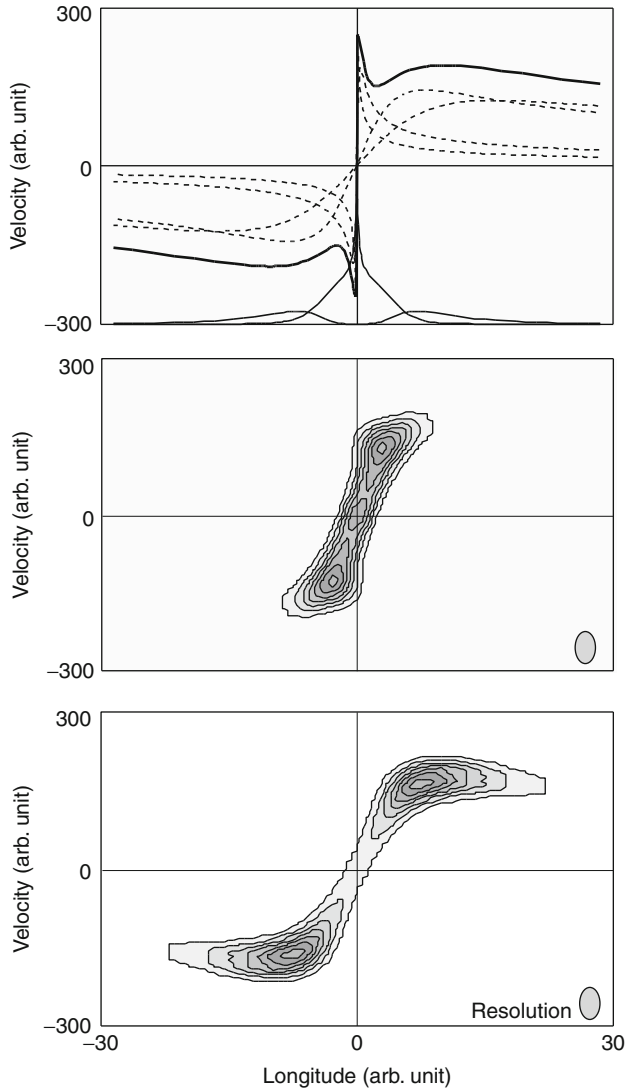
Rotation curve of a galaxy is defined as the trace of velocities on a position-velocity (PV) diagram along the major axis corrected for the inclination angle between the line of sight and the galaxy disk. There are several methods to derive the rotation velocities from the PV diagrams.

A widely used method is to trace intensity-weighted velocities, which are defined by

$$V_{\text{int}} = \frac{1}{\sin i} \int I(v)v dv / \int I(v)dv, \quad (19.14)$$

where $I(v)$ is the intensity profile at a given radius as a function of the radial velocity v corrected for the systemic velocity of the galaxy and i is the inclination angle. For convenience, the intensity-weighted velocity is often approximated by centroid velocity or peak-intensity velocity, which is close to each other. The centroid velocity is often obtained by tracing the values on the mean-velocity map of a disk galaxy, which is usually produced from a spectral data cube by taking the first moment. The mean velocities are obtained also by tracing values along the major axis of a mean-velocity map (moment 1 map) produced from a spectral data cube.

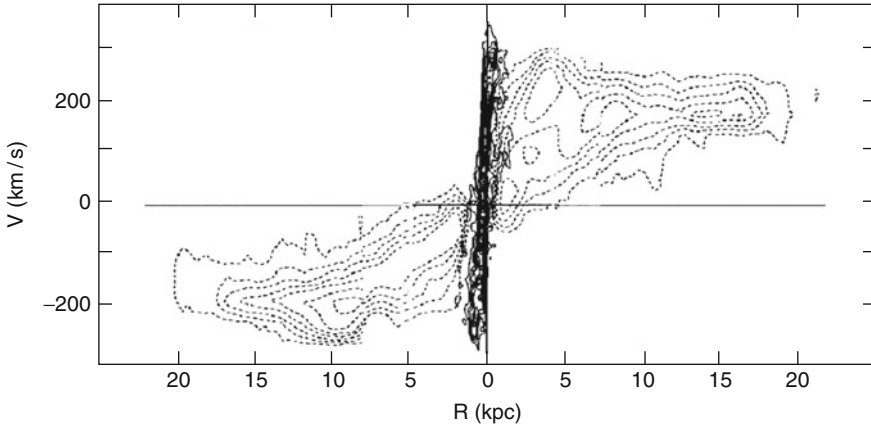
The obtained rotation velocity is reasonable for the outer disk. However, it is largely deviated from the true rotation speed in the innermost region, where the velocity structure is complicated. It should be remembered that the mean velocity near the nucleus gives always underestimated rotation velocity, because the finite resolution of observation inevitably results



■ Fig. 19-8

(Top) A model rotation curve comprising a massive core, bulge, disk, and halo. Distributions of the molecular (CO) and HI gases are given by *thin lines*. (Middle) Composed position-velocity diagram in CO, and (Bottom) HI

in zero value at the center by averaging plus and minus values in both sides of nucleus along the major axis. Hence, the derived rotation curve often starts from zero velocity in the center. But the nucleus is the place where the stars and gases are most violently moving, often nesting a black hole with the surrounding objects moving at high velocities close to the light speed.



■ Fig. 19-9

Position-velocity diagram along the major axis of the edge-on galaxy NGC 3079. Dashed and central full contours are from HI (Irwin et al. 1991) and CO ($J = 1-0$) (Sofue et al. 2001) line observations, respectively

2.3.4 Terminal-Velocity Method

This method makes use of the terminal velocity in a PV diagram along the major axis. The rotation velocity is derived by using the terminal velocity V_t :

$$V_{\text{rot}} = V_t / \sin i - (\Sigma_{\text{obs}}^2 + \Sigma_{\text{ISM}}^2)^{1/2}, \quad (19.15)$$

where Σ_{ISM} and Σ_{obs} are the velocity dispersion of the interstellar gas and the velocity resolution of observations, respectively. The interstellar velocity dispersion is of the order of $\Sigma_{\text{ISM}} \sim 5-10 \text{ km s}^{-1}$, while Σ_{obs} depends on the instruments.

Here, the terminal velocity is defined by a velocity at which the intensity becomes equal to $I_t = [(\eta I_{\text{max}})^2 + I_{\text{lc}}^2]^{1/2}$ on the observed PV diagram, where I_{max} and I_{lc} are the maximum intensity and intensity corresponding to the lowest contour level, respectively, and η is usually taken to be $\sim 0.2\%$ so that the 20% level of the intensity profile is traced. If the intensity is weak, the equation gives $I_t \simeq I_{\text{lc}}$ which approximately defines the loci along the lowest contour level.

2.3.5 Iteration Method

A more reliable method is to reproduce the observed position-velocity diagram by correcting the iteratively obtained rotation curves (Takamiya and Sofue 2002). The method comprises the following procedure. An initial rotation curve, RC0, is adopted from a PV diagram (PV0), obtained by any method as above (e.g., a peak-intensity method). Using this rotation curve and an observed radial distribution of intensity (emissivity) of the line used in the analysis, a PV diagram, PV1, is constructed. The difference between this calculated PV diagram and the original PV0 (e.g., the difference between peak-intensity velocities) is used to correct for the initial rotation curve to obtain a corrected rotation curve RC1. This RC is used to calculate another PV diagram PV2 using the observed intensity distribution, and to obtain the next iterated rotation curve, RC2 by correcting for the difference between PV2 and PV0. This iteration is repeated

until PV_i and PV_0 becomes identical, such that the summation of root mean square of the differences between PV_i and PV_0 becomes minimum and stable. RC_i is adopted as the most reliable rotation curve.

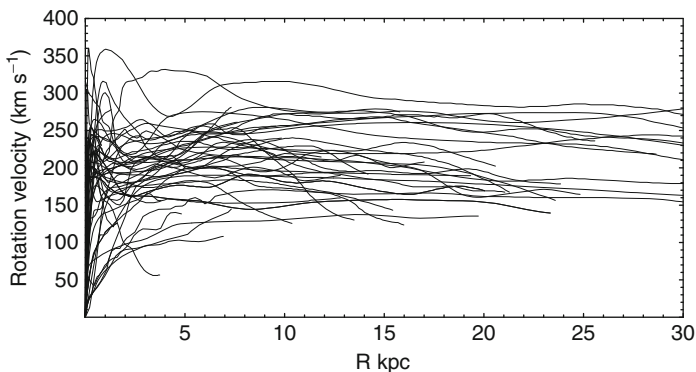
2.3.6 Three-Dimensional Cube Method for the Future

The methods so far described utilize only a portion of the kinematical data of a galaxy. The next-generation method to deduce galactic rotation would be to utilize three-dimensional spectral data from the entire galaxy. This may be particularly useful for radio line observations, because the lines are transparent at any places in the disk. In the iteration method position-velocity diagrams along the major axis are used. In the three-dimensional method, the entire spectral data will be employed, which consist of spectra at all two-dimensional grids in the galactic disk on the sky. Such data are usually recorded as a spectral cube, or intensities I at all points in the (x, y, v) space, as obtained by spectral imaging observations of radio lines such as CO and HI. Here, (x, y) is the position on the sky and v is radial velocity.

The reduction procedure would be similar to that for the iteration method: First, an approximate rotation curve is given, and a spectral cube is calculated from the curve based on the density distribution already derived from the projected intensity distribution in the galaxy on the sky. Next, the calculated cube (C) is compared with the observed cube (O) to find the difference. Then, the assumed rotation curve is iteratively corrected so that the difference between the O and C get minimized. The density distribution itself may be taken as another unknown profile to be measured during the iteration in addition to the rotation curve.

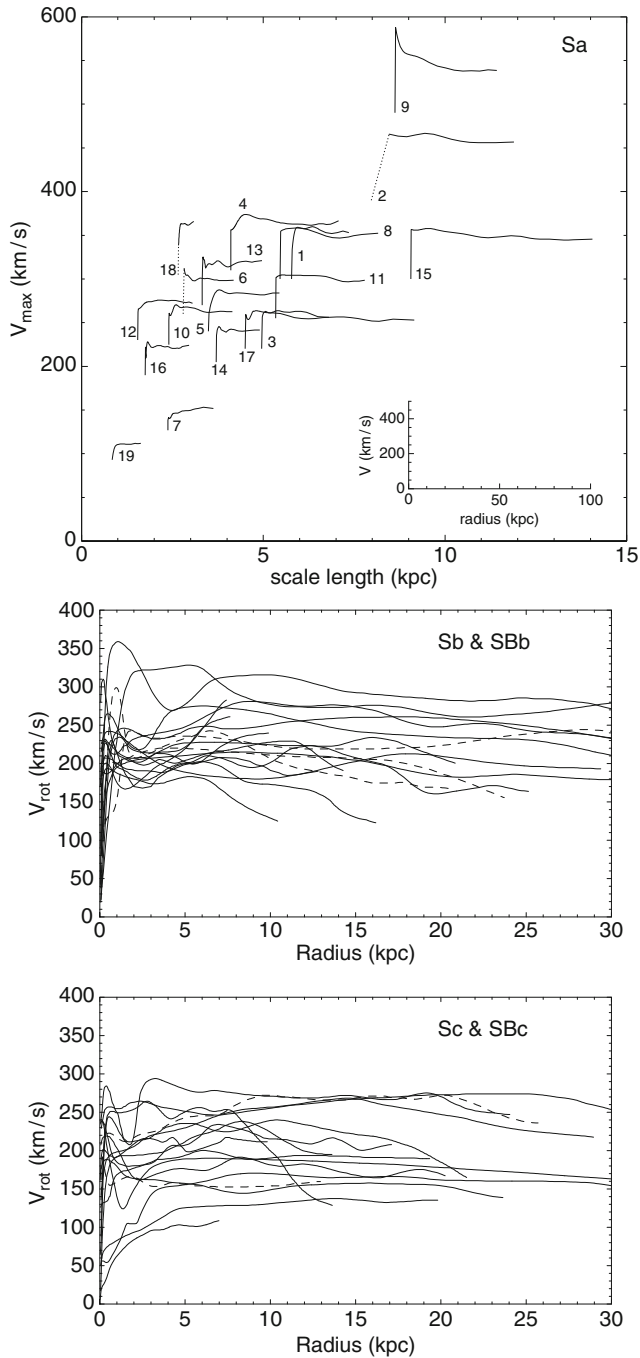
2.4 Rotation Curves of Spiral Galaxies

► [Figure 19-10](#) shows the rotation curves in nearby spiral galaxies, which have been obtained mainly by the terminal-velocity methods from optical, CO- and HI-line data (Sofue et al. 1999). ► [Figure 19-11](#) shows rotation curves for individual galaxies types from Sa to Sc. It is



■ Fig. 19-10

Rotation curves of spiral galaxies obtained by combining CO data for the central regions, optical for disks, and HI for outer disk and halo (Sofue et al. 1999)



■ Fig. 19-11

(Top) Rotation curves of early type spiral galaxies (Sa) listed on the maximum velocity-scale radius plane (Noordermeer et al. 2007). (Middle) Rotation curves for Sb galaxies (*full lines*) and barred SBb galaxies (*dashed lines*). (Lower panel) Same, but for Sc and SBc galaxies (Sofue et al. 1999)

remarkable that the form, but not amplitude, of the disk and halo rotation curves is similar to each other for different morphologies from Sa to Sc. This suggests that the form of the gravitational potential in the disk and halo is not strongly dependent on the galaxy type. A moderate correlation is found between total luminosity and maximum rotation velocity, which is known as the Tully–Fisher relation (Tully and Fisher 1977; Mathewson et al. 1992; Mathewson and Ford 1996). Less luminous galaxies tend to show increasing outer rotation curve, while most massive galaxies have flat or slightly declining rotation in the outermost part (Persic et al. 1996).

2.4.1 Sa, Sb, Sc Galaxies

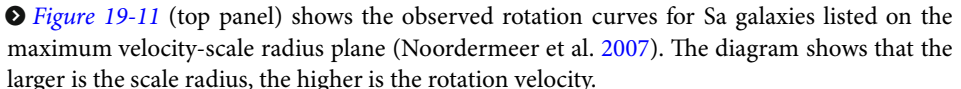
The form of central rotation curves depends on the total mass and galaxy types (Sofue et al. 1999). Massive galaxies of Sa and Sb types show a steeper rise and higher central velocities within a few hundred parsecs of the nucleus compared to less massive Sc galaxies and dwarfs (Noordermeer et al. 2007). Dwarf galaxies generally show a gentle central rise. This is related to the bulge-to-disk mass ratio: The smaller is the ratio, and therefore, the later is the type, the weaker is the central rise (Casertano and van Gorkom 1991; Rubin et al. 1985).

For massive Sb galaxies, the rotation maximum appears at a radius of 5 or 6 kpc, which is about twice the scale radius of the disk. Beyond the maximum, the rotation curve is usually flat, merging with the flat part due to the massive dark halo. Fluctuations of a few tens of km s^{-1} due to non-axisymmetric structures such as spiral arms and bars are superposed as velocity ripples. The fluctuations of order of 50 km s^{-1} are often superposed for barred galaxies, which indicate noncircular motions in the oval potential.

There have been several attempts to represent the observed rotation curves by simple functions (Courteau 1997; Persic et al. 1996; Roscoe 1999). Persic et al. fit the curves by a formula, which is a function of total luminosity and radius, comprising both disk and halo components. Both the forms and amplitudes are functions of the luminosity, and the outer gradient of the RC is a decreasing function of luminosity.

Rotation curves of Sa and Sb galaxies, including the Milky Way, can be summarized to have:

- (a) High-velocity rotation or dispersion in the nucleus due to a high-density core and massive black hole with declining velocity to a minimum at $R \sim$ a few tens parsecs
- (b) Steep rise of RC within the central 100 pc due to the bulge
- (c) Maximum at radius of a few hundred parsecs due to the bulge, followed by a decline to a minimum at 1–2 kpc; then
- (d) Gradual rise to the maximum at 5–7 kpc due to the disk and
- (e) Nearly flat outer rotation due to the dark halo up to $R \sim 20\text{--}30$ kpc

Earlier type galaxies have higher maximum rotation velocities than later type galaxies: Sa galaxies have maximum velocities of around 300 km s^{-1} and Sb 220 (Rubin et al. 1985).  **Figure 19-11** (top panel) shows the observed rotation curves for Sa galaxies listed on the maximum velocity-scale radius plane (Noordermeer et al. 2007). The diagram shows that the larger is the scale radius, the higher is the rotation velocity.

Sc galaxies have lower maximum velocities than Sa and Sb, ranging from ≤ 100 to $\sim 200 \text{ km s}^{-1}$ with the median value of 175 km s^{-1} (Rubin et al. 1985). Massive Sc galaxies show a steep nuclear rise similar to Sb's, while less massive Sc galaxies have gentler rise. They also have a flat rotation to their outer edges. Less luminous (lower surface brightness) Sc galaxies have a gentle

central rise of rotation velocity, which monotonically increases till the outer edge. This behavior is similar to rotation curves of dwarf galaxies.

2.4.2 Barred Galaxies

Barred spiral galaxies constitute a considerable fraction of all disk galaxies. Large-scale rotation properties of SBb and SBc galaxies are generally similar to those of non-barred galaxies of Sb and Sc types. However, their kinematics is more complicated due to the noncircular streaming motion by the oval potential, which results in skewed velocity fields and ripples on the rotation curves. Large velocity variation arises from the barred potential of several kiloparsecs length on the order of $\pm \sim 50\text{--}100 \text{ km s}^{-1}$ (e.g., Kuno et al. 2000). Simulations of PV diagrams for edge-on barred galaxies show many large velocity fluctuations superposed on the flat rotation curve (Bureau and Athanassoula 1999; Weiner and Sellwood 1999). However, distinguishing the existence of a bar and quantifying it are not uniquely done from such limited edge-on information. For more quantitative results, two-dimensional velocity analyses are necessary (Wozniak and Pfenniger 1997). However, in contrast to the basic mass structures as the bulge, disk, and halo, determination of the mass distribution in the bar from observed data is still not straightforward because of its large number of parameters to be fixed, which are the bar's three axial lengths, major axis direction, density amplitude, and mass distribution. Also, its dynamical connection to the disk and bulge is a subject for further investigation.

2.4.3 Dwarf Galaxies

The Large Magellanic Cloud is the nearest dwarf galaxy, which shows a flat rotation curve from the dynamical center to the outer edge at $R \sim 5 \text{ kpc}$ at a velocity $\sim 100 \text{ km s}^{-1}$ (Sofue 2000). This galaxy reveals a particular kinematical property: The dynamical center is significantly displaced from the optical bar center, indicating the existence of a massive bulge-like density component that is not visible. The Small Magellanic Cloud shows complex velocity field spitted into two velocity components. The velocity data are not appropriate to discuss the mass in this very disturbed galaxy.

Dwarf and low surface brightness galaxies show slow rotation at $\leq 100 \text{ km s}^{-1}$ with monotonically rising rotation curves their last measured points (Blais-Ouellette et al. 2001; Carignan 1985; de Blok 2005; Swaters et al. 2009). Due to the low luminosity, the mass-to-luminosity ratio is usually higher than that for normal spiral, and the dark matter fraction is much higher in dwarf galaxy than in spirals (Carignan 1985; Jobin and Carignan 1990).

2.4.4 Irregular Galaxies

The peculiar morphology of irregular galaxies is mostly produced by gravitational interaction with the companion and/or an encountered. When galaxies gravitationally interact, they tidally distort each other, and produce the peculiar morphology that had until recently defied classification. Toomre and Toomre (1972) for the first time computed tidal interactions to simulate some typical distorted galaxies.

The starburst dwarf galaxy NGC 3034 (M82) shows an exceptionally peculiar rotation property (Burbidge et al. 1964; Burbidge and Burbidge 1975). It has a normal nuclear rise and rotation velocities which have a Keplerian decline beyond the nuclear peak. This may arise from a tidal truncation of the disk and/or halo by an encounter with M81 (Sofue 1998).

3 Galactic Mass Distribution

3.1 Approximate Mass Distribution

3.1.1 Flat Rotation and Isothermal Mass Distribution

It has been shown that the rotation curve of the Galaxy as well as those of most spiral galaxies is nearly flat, indicating that the rotation velocity V is approximately constant in a galaxy. Before describing the detailed mass models, an approximate mass profile in the Galaxy may be estimated from the flat characteristics of the entire rotation curve. Simply assuming a spherical distribution, the mass involved within a radius R is approximated by

$$M(R) \sim \frac{RV(R)^2}{G}. \quad (19.16)$$

If the rotation velocity V is constant, remembering that $M(R) = 2\pi \int_0^R \Sigma r dr$ and $M(R) = 4\pi \int_0^R \rho r^2 dr$, the surface mass density (SMD) is approximately given by

$$\Sigma \propto R^{-1}, \quad (19.17)$$

and volume mass density behaves similarly to an isothermal gas sphere as

$$\rho \propto R^{-2}. \quad (19.18)$$

These are very rough and the first-order mass distributions in a galaxy having flat rotation curve.

3.1.2 Vertical Mass Distribution Near the Sun

If the local stellar distribution is assumed to be a flat disk around the galactic plane, an approximate density profile perpendicular to the galactic plane in the solar vicinity can be calculated without considering the galactic rotation. Since motions of stars and gas are deviating from circular orbits, the deviation acts as the pressure, which yields the disk thickness. The local distribution of stars and gas perpendicular to the galactic plane is approximated by static equilibrium between the z -directional gravity (perpendicular to the disk) and the velocity dispersion in the z direction v_z . By denoting the equivalent pressure of the matter in the disk as $p = \rho v_z^2$, the equilibrium in the z direction can be written as

$$\frac{dp}{dz} = \frac{d(\rho v_z^2)}{dz} = g_z \rho, \quad (19.19)$$

where g_z and ρ are the z -directional gravitational acceleration and the density of the matter (stars and gas). If the galactic disk is assumed to be a flat plane of infinite surface area, the

gravity can be approximated as

$$g_z \sim -\alpha z, \quad (19.20)$$

and therefore,

$$\frac{d\rho}{dz} \sim -\frac{\alpha z}{v_z^2} \rho. \quad (19.21)$$

This equation can be solved if the velocity dispersion is constant with z :

$$\rho = \rho_0 \exp\left(-\frac{z^2}{z_0^2}\right), \quad (19.22)$$

where ρ_0 and z_0 are constants depending on the species. If the Poisson equation is solved for a self-gravitating flat disk, a slightly different profile (Spitzer 1942) is obtained as

$$\rho = \rho_0 \operatorname{sech}^2\left(\frac{z}{z_0}\right). \quad (19.23)$$

Here, z_0 , is the z -directional scale height representing the typical thickness of the disk, and is related to the velocity dispersion as

$$z_0 = \frac{v_z}{\sqrt{\alpha}} \sim \sqrt{\frac{k}{\alpha m_H}} T_{\text{gas}}. \quad (19.24)$$

In **Fig. 19-12** the two results from **(19.22)** and **(19.23)** are compared. **Table 19-1** summarizes the typical values of v_z and the disk scale thickness z_0 for various species. Population I objects are distributed in a disk of thickness $2z_0 \sim 100$ pc, and Population II stars in a disk with thickness $2z_0 \sim 300$ pc. See **Table 19-3** for more recent observed values.

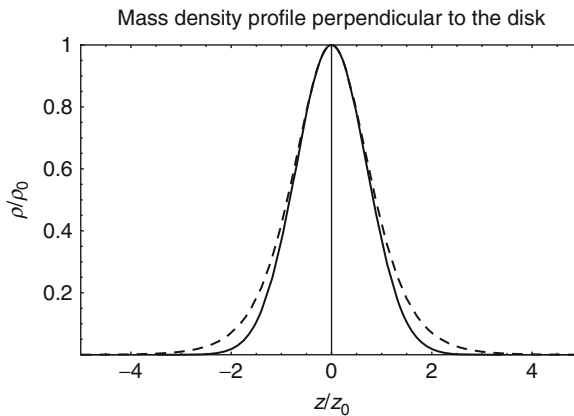


Fig. 19-12

Mass density profile perpendicular to the local disk. The solid curve represents the Gaussian profile for a hydrostatic disk, and the dashed line is for a self-gravitating disk (Spitzer 1942)

■ **Table 19-1**

Velocity dispersion in the z direction and scale thickness of the disk

Component	v_z (km s ⁻¹)	Thickness, $2z_0$ (pc)
Molecular clouds	5	60
OB stars	5	60
HI gas	10	100
Disk stars	20–30	300
Hot gas (10 ⁶ K)	100 ($\sim \sqrt{\frac{k}{m_H T}}$)	~ 2 kpc

3.2 Decomposition Methods

The observed rotation curve is used to derive the mass distribution in the Galaxy. There are two ways to derive the mass distribution from the rotation curve.

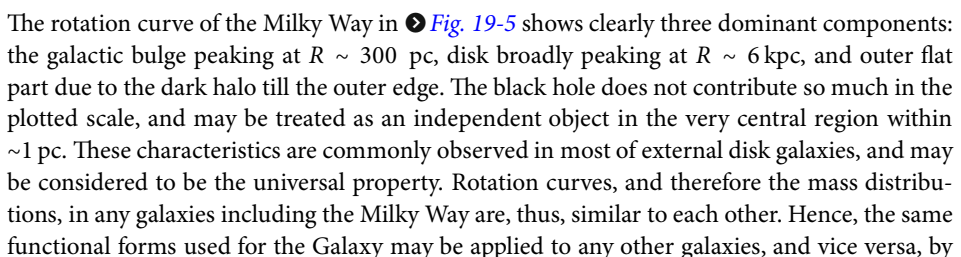
Decomposition method: One way is to fit the observed rotation curve by a calculated one assuming that the galaxy is composed of several components, such as a central bulge, disk, and a dark halo. The total rotation velocity is given by summing up the squares of individual corresponding velocities:

$$V(R) = \sqrt{\sum V_i^2} \simeq \sqrt{V_b(R)^2 + V_d(R)^2 + V_h(R)^2}. \quad (19.25)$$

If the resolution in the central region is sufficiently high, a more central component such as a black hole such as $V_{\text{BH}}(R)^2$ may be added to the right-hand side. Here, V_i indicates circular velocity corresponding to the i -th component alone, and suffices BH, b, d, and h denote a black hole, bulge, disk, and halo components, respectively. For the fitting, the parameters such as the mass and scale radius of individual components are adjusted so that the residual between the calculation and observation gets minimized. The fitting is usually started from the inner component. First, the innermost steep rise and peak of the rotation curve is fitted by a bulge component; second, the gradual rise and flat part is fitted by the disk, and finally, the residual outskirts are fitted by a dark halo. Alternatively, these fittings may be done at the same time by giving approximate parameters, which are adjusted iteratively until the entire curve is best fitted.

Direct method: Another way is to calculate the mass distribution directly from the rotation curve. This method gives more accurate mass distributions without being affected by the assumed components and their functional forms.

3.3 Decomposition into Bulge, Disk, and Halo

The rotation curve of the Milky Way in  Fig. 19-5 shows clearly three dominant components: the galactic bulge peaking at $R \sim 300$ pc, disk broadly peaking at $R \sim 6$ kpc, and outer flat part due to the dark halo till the outer edge. The black hole does not contribute so much in the plotted scale, and may be treated as an independent object in the very central region within ~ 1 pc. These characteristics are commonly observed in most of external disk galaxies, and may be considered to be the universal property. Rotation curves, and therefore the mass distributions, in any galaxies including the Milky Way are, thus, similar to each other. Hence, the same functional forms used for the Galaxy may be applied to any other galaxies, and vice versa, by

modifying the parameters such as the masses and scale radii of individual components. Decomposition of a rotation curve into several mass components such as a bulge, disk, and dark halo has been extensively applied to observed data (Bosma 1981a; Kent 1986; Sofue 1996).

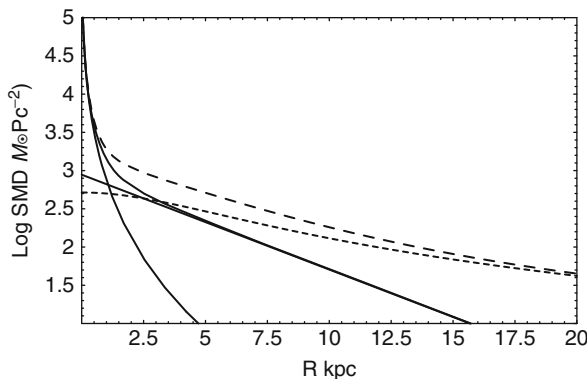
There are various functional forms to represent the mass components. The commonly employed mass profiles are:

1. *de Vaucouleurs bulge + exponential disk*: It is well established that the luminosity profile of the spheroidal bulge component in galaxies is represented by the de Vaucouleurs ($e^{-(r/r_c)^{1/4}}$; 1953, 1958) or Sérsic ($e^{-(r/r_c)^n}$; 1968) law. The disk component is represented by the exponential law (e^{-r/r_c} ($n = 1$); Freeman 1970). For the halo, either isothermal, Navarro–Frenk–White (NFW), or Burkert profile is adopted.
2. *Miyamoto–Nagai potential*: Another way is to assume a modified Plummer-type potential for each component, and the most convenient profile is given by the Miyamoto–Nagai (1975) potential.

After fitting the rotation curve by the bulge, exponential disk, and a dark halo, smaller structures superposed on the rotation curves are discussed as due to more local structures such as spiral arms and/or bar.

3.3.1 de Vaucouleurs Bulge

The most conventional method among the decomposition methods is to use functional forms similar to luminosity profiles. The inner region of the galaxy is assumed to be composed of two luminous components, which are a bulge and a disk. The mass-to-luminosity ratio within each component is assumed to be constant, so that the mass density distribution has the same profile. The bulge is assumed to have a spherically symmetric mass distribution, whose surface mass density obeys the de Vaucouleurs law, as shown in ► Fig. 19-13.



■ Fig. 19-13

Surface mass density (SMD) distributions in the Galaxy. The *thin curve* shows the de Vaucouleurs bulge, and the *straight line* the exponential disk. Their sum is also shown. The *dashed line* is the dark halo, and *long-dashed line* indicates the total SMD. High density at the center due to the bulge reaches $6.8 \times 10^6 M_{\odot} \text{pc}^{-2}$

The de Vaucouleurs (1958) law for the surface brightness profile as a function of the projected radius R is expressed by

$$\log \beta = -\gamma(\alpha^{1/4} - 1), \quad (19.26)$$

with $\gamma = 3.3308$. Here, $\beta = B_b(R)/B_{be}$, $\alpha = R/R_b$, and $B_b(R)$ is the brightness distribution normalized by B_{be} , which is the brightness at radius R_b . The same de Vaucouleurs profile for the surface mass density is adopted as

$$\Sigma_b(R) = \lambda_b B_b(R) = \Sigma_{bc} \exp \left[-\kappa \left(\left(\frac{R}{R_b} \right)^{1/4} - 1 \right) \right] \quad (19.27)$$

with $\Sigma_{bc} = 2142.0 \Sigma_{be}$ for $\kappa = \gamma \ln 10 = 7.6695$. Here, λ_b is the mass-to-luminosity ratio, which is assumed to be constant within a bulge. Equations 19.26 and 19.27 have a particular characteristics: The central value at $r = 0$ is finite, and the function decreases very steeply with radius near the center. However, the gradient gets milder as radius increases, and the SMD decreases very slowly at large radius, forming an extended outskirts. The function well represents the brightness distribution in spheroidal components and elliptical galaxies that have a strong concentration toward the center with finite amplitude, while the outskirts extends widely.

The total mass is calculated by

$$M_{bt} = 2\pi \int_0^\infty r \Sigma_b(r) dr = \eta R_b^2 \Sigma_{bc}, \quad (19.28)$$

where $\eta = 22.665$ is a dimensionless constant. By definition, a half of the total projected mass (luminosity) is equal to that inside a cylinder of radius R_b .

In order to describe the bulge component in the Galaxy, it is often assumed that the bulge is spherical having the de Vaucouleurs type profile. The volume mass density $\rho(R)$ at radius r for a spherical bulge is calculated by using the surface density as

$$\rho(R) = \frac{1}{\pi} \int_R^\infty \frac{d\Sigma_b(x)}{dx} \frac{1}{\sqrt{x^2 - R^2}} dx. \quad (19.29)$$

Since the mass distribution is assumed to be spherical, the circular velocity is calculated from the total mass enclosed within a sphere of radius R :

$$V_b(R) = \sqrt{\frac{GM_b(R)}{R}}. \quad (19.30)$$

The velocity approaches the Keplerian-law value at radii sufficiently greater than the scale radius. The shape of the rotation curve is similar to each other for varying total mass and scale radius. For a given scale radius, the peak velocity varies proportionally to a square root of the mass. For a fixed total mass, the peak-velocity position moves inversely proportionally to the scale radius along a Keplerian line.

Decomposition of rotation curves by the $e^{-(R/r_e)^{1/4}}$ law surface mass profiles have been extensively applied to spheroidal components of late type galaxies (Noordermeer 2008; Noordermeer et al. 2007). Sérsic (1968) has proposed a more general form $e^{-(R/r_e)^n}$ for spheroidal luminosity distributions. The $e^{-R^{1/4}}$ law was fully discussed in relation to its dynamical relation to the galactic structure based on the more general profile (Ciotti 1991; Trujillo et al. 2002).

Figure 19-6 shows the calculated rotation curve for the de Vaucouleurs bulge model. The result shows a reasonable fit of the inner rotation curve. It reproduces the central steep rise and

sharp peak at $R = 300$ pc, which is due to the high-density mass concentration in the central region.

The best-fit total mass of the bulge of our Galaxy is as large as $1.8 \times 10^{10} M_{\odot}$ and the scale radius is as compact as $R_b = 0.5$ kpc with the accuracy of about 5%. As [Fig. 19-13](#) indicates, the surface mass density in the central 0.5 kpc is dominated by the bulge component, and nuclear surface density reaches a value as high as $6.8 \times 10^6 M_{\odot} \text{pc}^{-2}$. The total projected mass included in the central 500 pc (scale radius) is $9.0 \times 10^9 M_{\odot}$. The total spheroidal mass integrated in a sphere of R_b is 0.39 times the total bulge mass, which is for the present case $7.0 \times 10^9 M_{\odot}$.

3.3.2 M/L Ratio in Bulge

The surface mass distribution in [Fig. 19-13](#) may be compared with the K band surface brightness profile for the inner Galaxy at $|b| < 10^\circ$ as presented by Kent et al. (1991) and Kent (1992). The K band luminosity profile for the central 1 kpc is expressed as

$$v_K = 1.04 \times 10^6 (r/0.482 \text{pc})^{-1.85} L_{\odot} \text{pc}^{-3}. \quad (19.31)$$

This expression approximates the density profile for an isothermal sphere with a constant mass-to-luminosity ratio, for which the power-law index is -2 . Since the functional form is different, it cannot be compared directly with the de Vaucouleurs profile. However, it may be worthy to compare the integrated luminosity within a radius with the corresponding mass. The integrated luminosity within a sphere of radius 0.5 kpc from [\(19.31\)](#) is calculated to be $3.74 \times 10^9 L_{\odot}$. Thus, a mass-to-luminosity ratio for the bulge within a sphere of the scale radius, $R_b = 0.5$ kpc, is obtained to be $M/L(\text{bulge in } 0.5 \text{ kpc}) \simeq 7.1 M_{\odot}/L_{\odot}$.

As [Fig. 19-13](#) indicates, the volume density increases rapidly toward the Galactic Center, approaching an infinite value. Note that the SMD at the center is finite as represented by [\(19.27\)](#), whereas the volume density diverges to infinity at the center as in [\(19.29\)](#). The central mass within 1 pc is estimated to be as high as several $10^6 M_{\odot}$.

3.3.3 Exponential Disk

The galactic disk is represented by an exponential disk (Freeman 1970). The surface mass density is expressed as

$$\Sigma_d(R) = \Sigma_{dc} \exp(-R/R_d), \quad (19.32)$$

where Σ_{dc} is the central value, R_d is the scale radius. The total mass of the exponential disk is given by $M_{\text{disk}} = 2\pi \Sigma_{dc} R_d^2$. The rotation curve for a thin exponential disk without perturbation, e.g., $\Delta = 0$, is expressed by (Binney and Tremaine 1987; Freeman 1970).

$$V_d = \sqrt{R \frac{\partial \Phi}{\partial R}} = \sqrt{4\pi G \Sigma_0 R_d y^2 [I_0(y)K_0(y) - I_1(y)K_1(y)]}, \quad (19.33)$$

where $y = R/(2R_d)$, and I_i and K_i are the modified Bessel functions.

If the rotation curve is affected by additional masses Δ due to arms, rings, bar, and/or interstellar gas, the surface mass density (SMD) is represented as

$$\Sigma_d(R) = \Sigma_{dc} \exp(-R/R_d) + \Delta, \quad (19.34)$$

The gravitational force $f(R)$ acting on a point at galactocentric distance $x = R$ is then directly calculated by integrating the x directional component of force due to a mass element $\Sigma'_d(x)dx dy$ in the Cartesian coordinates (x, y) :

$$f(R) = G \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\Sigma_d(x)(R-x)}{s^3} dx dy, \quad (19.35)$$

where $s = \sqrt{(R-x)^2 + y^2}$ is the distance between the mass element and the point. The rotation velocity is given by

$$V_d(R) = \sqrt{fR}. \quad (19.36)$$

3.3.4 Dark Halo

Fitting by a dark halo model within ~ 20 kpc is crude not only because of the large scatter and errors of observed rotation velocities in the outer disk, but also for the weaker response of the rotation curve to the halo models. The data may be fitted by an isothermal halo model with the asymptotic velocity of 200 km s^{-1} at infinity and a scale radius of $h = 10$ kpc. Reasonable fit is also obtained with the same scale radius for the NFW and Burkert profiles, for which the rotation velocity at 20 kpc is approximately 200 km s^{-1} . The result is weakly dependent on the adopted model. The dark halo models are described in detail in the next section, the rotation curve is combined with radial velocities of objects surrounding the Galactic halo such as globular clusters, satellite galaxies, and member galaxies in the Local Group.

3.4 Galactic Mass Parameters

Using the rotation curve of the Galaxy as in [Fig. 19-5](#) and adopting the classical decomposition method, the parameters are obtained for individual mass components of the bulge, disk, and dark halo. The functional form of the bulge was so adopted that the surface mass density is represented by the de Vaucouleurs law. The disk was approximated by an exponential disk, and the halo by an isothermal sphere. The observed characteristics are well fitted by superposition of these components. The central steep rise and the high rotation peak at $R = 300$ pc is quite well reproduced by the de Vaucouleurs bulge of half-mass scale radius $R_b = 0.5$ kpc. The broad maximum at around $R \sim 6$ kpc was fitted by the exponential disk, and the flat outer part by a usual dark halo. [Table 19-2](#) lists the fitting parameters for individual mass components. Since the used data in [Fig. 19-5](#) were compiled from different observations, their errors are not uniform, and only eye-estimated values are given, which were evaluated after trial and error of fitting to the observed points.

The local values of the surface mass and volume densities in the solar vicinity calculated for these parameters are also shown in [Table 19-3](#). The volume density of the disk has been calculated by $\rho_d = \Sigma_d/(2z_0)$ with z_0 being the scale height at $R = R_0$, when the disk scale profile is approximated by $\rho_d(R_0, z) = \rho_{d0}(R_0)\text{sech}(z/z_0)$. For the local galactic disk, two values are adopted: $z_0 = 144 \pm 10$ pc for late type stars based on the Hipparcos star catalogue (Kong and Zhu 2008) and 247 pc from Kent et al. (1991). The local volume density by the bulge is four orders of magnitudes smaller than the disk component, and the halo density is two orders of magnitudes smaller. However, the surface mass densities as projected on the Galactic plane are not negligible. The bulge contributes to 1.6% of the disk value, or the stars in the direction of

■ Table 19-2
Parameters for Galactic mass components^a

Component	Parameter	Value ^b
Massive black hole ^c	Mass	$M_{\text{BH}} = 3.7 \times 10^6 M_{\odot}$
Bulge	Mass	$M_{\text{b}} = 1.80 \times 10^{10} M_{\odot}$
	Half-mass scale radius	$R_{\text{b}} = 0.5 \text{ kpc}$
	SMD at R_{b}	$\Sigma_{\text{be}} = 3.2 \times 10^3 M_{\odot} \text{pc}^{-2}$
	Center SMD	$\Sigma_{\text{bc}} = 6.8 \times 10^6 M_{\odot} \text{pc}^{-2}$
	Center volume density	$\rho_{\text{bc}} = \infty$
Disk	Mass	$M_{\text{d}} = 6.5 \times 10^{10}$
	Scale radius	$R_{\text{d}} = 3.5 \text{ kpc}$
	Center SMD	$\Sigma_{\text{dc}} = 8.44 \times 10^2 M_{\odot} \text{pc}^{-2}$
	Center volume density	$\rho_{\text{dc}} = 8 M_{\odot} \text{pc}^{-3}$
Bulge and disks	Total mass	$M_{\text{b+d}} = 8.3 \times 10^{10} M_{\odot}$
Dark halo (Isothermal sphere)	Mass in $r \leq 10 \text{ kpc}$ sphere	$M_{\text{h}}(\leq 10 \text{ kpc}) \approx 1.5 \times 10^{10} M_{\odot}$
	Mass in $r \leq 20 \text{ kpc}$ sphere ^d	$M_{\text{h}}(\leq 20 \text{ kpc}) \approx 7.1 \times 10^{10} M_{\odot}$
	Core radius	$h \approx 12 \text{ kpc}$
	Central SMD in $ z < 10 \text{ kpc}$	$\Sigma_{\text{hc}} \approx 4.4 \times 10^2 M_{\odot} \text{pc}^{-2}$
	Central volume density	$\rho_{\text{hc}} \approx 5.1 \times 10^{-3} M_{\odot} \text{pc}^{-3}$
	Circular velocity at infinity	$V_{\infty} \approx 200 \text{ km s}^{-1}$
Total galactic mass	Mass in $r \leq 20 \text{ kpc}$ sphere	$M_{\text{total}}(\leq 20 \text{ kpc}) \approx 1.5 \times 10^{11} M_{\odot}$
	Mass in $r \leq 100 \text{ kpc}$ NFW sphere	$M_{\text{total}}(\leq 100 \text{ kpc}) \approx 3 \times 10^{11} M_{\odot}$

^aSofue et al. (2009)

^bUncertainty is $\sim 10\%$ for bulge and disk, and $\sim 20\%$ for halo

^cGenzel et al. (2000)

^dMass within 20 kpc is weakly dependent on the halo models, roughly equal to those for NFW and Burkert models. At larger distances beyond 30 kpc, it varies among the halo models (Sofue 2009)

the galactic pole would include about 2% bulge stars, given the de Vaucouleurs density profile. The surface mass density of the dark halo integrated at heights of $-10 < z < 10 \text{ kpc}$ exceeds the disk value by several times.

3.5 Miyamoto–Nagai Potential

Although the deconvolution using the de Vaucouleurs and exponential disk well represents the observations, the mass models are not necessarily self-consistent in the sense that the model mass profiles are the solutions of the Poisson equation. One of the convenient methods to represent the Galaxy's mass distribution by a self-consistent dynamical solution is to use superposition of multiple Plummer-type potentials.

The mass distribution $\rho(R, z)$ and the gravitational potential $\Phi(R, z)$ are related by Poisson's equation:

$$\Delta\Phi = 4\pi\rho(R, z), \quad (19.37)$$

■ Table 19-3

Local values near the Sun at $R = R_0$ (8.0 kpc)

	Components	Local values
Surface mass density	Bulge (de Vaucouleurs)	$\Sigma_b^\odot = 1.48M_\odot\text{pc}^{-2}$
	Disk (exponential; including gas)	$\Sigma_d^\odot = 87.5M_\odot\text{pc}^{-2}$
	Interstellar gas (HI + H ₂)	$\Sigma_{\text{gas}}^\odot = 5.0M_\odot\text{pc}^{-2}$
	Bulge+Disk	$\Sigma_{\text{bd}}^\odot = 89M_\odot\text{pc}^{-2}$
	Dark halo (isothermal, $ z < 10$ kpc)	$\Sigma_{\text{halo}}^\odot = 3.2 \times 10^2 M_\odot\text{pc}^{-2}$
	Total (bulge+disk+halo)	$\Sigma_{\text{total}}^\odot = 4.2 \times 10^2 M_\odot\text{pc}^{-2}$
Volume mass density	Bulge	$\rho_b^\odot = 1.3 \times 10^{-4} M_\odot\text{pc}^{-3}$
	Disk ^a for $z_0 = 144$ pc	$\rho_d^\odot = 0.30M_\odot\text{pc}^{-3}$ ($= \Sigma_d^\odot/2z_0$)
	Disk for $z_0 = 247$ pc	$\rho_d^\odot = 0.18M_\odot\text{pc}^{-3}$
	Disk Oort's value	$\rho_d^\odot = 0.15M_\odot\text{pc}^{-3}$
	Interstellar gas (included in disk)	$\rho_{\text{gas}}^\odot = 0.05M_\odot\text{pc}^{-3}$
	Bulge+Disk	$\rho_{\text{bd}}^\odot = 0.18 - 0.3M_\odot\text{pc}^{-3}$
	Dark halo	$\rho_{\text{halo}}^\odot \sim 3.5 \times 10^{-3} M_\odot\text{pc}^{-3}$
	Total (bulge+disk+halo)	$\rho_{\text{total}}^\odot = 0.2 - 0.3M_\odot\text{pc}^{-3}$
Total mass in solar sphere ^b ($r \leq R_0$)	Bulge	$M_b^\odot = 1.75 \times 10^{10} M_\odot$
	Disk	$M_d^\odot = 4.33 \times 10^{10} M_\odot$
	Bulge+Disk	$M_{\text{bd}}^\odot = 6.08 \times 10^{10} M_\odot$
	Dark halo	$M_{\text{dh}}^\odot \sim 8.7 \times 10^9 M_\odot$
	Total (bulge+disk+halo)	$M_{\text{total}}^\odot = 7.3 \times 10^{10} M_\odot$
	Representative mass $M_0 = R_0 V_0^2/G$	$M_0 = 7.44 \times 10^{10} M_\odot$

^aFor the scale height, $z_0 = 247$ pc (Kent et al. 1991) and 144 pc (Kong and Zhu 2008) are adopted

^bThe "Solar sphere" is a sphere of radius $R_0 = 8$ kpc centered on the Galactic Center

Let us recall that the potential for a point mass is given by

$$\Phi = \frac{-GM}{r} = \frac{-GM}{\sqrt{R^2 + z^2}}, \quad (19.38)$$

with $r = \sqrt{R^2 + z^2}$ being the distance from the center. An extended spherical mass is often described by Plummer's law:

$$\Phi = \frac{-GM}{\sqrt{r^2 + b^2}} = \frac{-GM}{\sqrt{R^2 + z^2 + b^2}}. \quad (19.39)$$

Here, b is a constant representing the scale radius of the sphere.

The most convenient Plummer-type formula, which describes the potential and realistic mass distribution in the Galaxy has been obtained by Miyamoto and Nagai (1975). The potential is a modified one from (19.39) for an axisymmetric spheroid. A galaxy is represented by superposition of several mass components in the same functional form as

$$\Phi = \sum_{i=1}^n \Phi_i = \sum_{i=1}^n \frac{-GM_i}{\sqrt{R^2 + (a_i + \sqrt{z^2 + b_i^2})^2}}, \quad (19.40)$$

where, a_i and b_i are constants representing the scale radius and scale height of the i -th spheroidal component. The rotation velocity in the galactic plane at $z = 0$ is given by

$$V_{\text{rot}}(R) = \sqrt{\sum_{i=1}^n R \frac{\partial \Phi_i}{\partial R}} = R \sqrt{\sum_{i=1}^n \frac{GM_i}{[R^2 + (a_i + b_i)^2]^{3/2}}}. \quad (19.41)$$

The mass distribution is calculated from Poisson's equation (● 19.37):

$$\rho(R, z) = \frac{1}{4\pi} \sum_{i=1}^n M_i \frac{a_i R^2 + [a_i + 3(z^2 + b_i^2)^{1/2}][a_i + (z^2 + b_i^2)^{1/2}]^2}{\{R^2 + [a_i + (z^2 + b_i^2)^{1/2}]^2\}^{5/2} (z^2 + b_i^2)^{3/2}}. \quad (19.42)$$

● *Figure 19-14* shows the meridional distribution of volume mass density calculated for a model Galaxy composed of two components of a bulge and disk with the parameters as given in ● *Table 19-4* (Miyamoto and Nagai 1975). This model approximately reproduces the rotation curve and the Oort's (1965) value of the local mass density of $0.15 M_{\odot} \text{pc}^{-3}$ at $R = 10$ kpc. Note that the dark halo was not taken into account, but instead the mass ($2.5 \times 10^{11} M_{\odot}$) and scale radius (7.5 kpc) were taken to be larger than the present-day values of $\sim 10^{11} M_{\odot}$ and ~ 3.5 kpc, in order to mimic the flat part of the rotation curve. Today, as discussed in the next section, the outer flat rotation is well understood as due to the dark halo. Nevertheless, the MN potential is often used, for its analytical form, to represent a galaxy by modifying the parameters and adding the halo and central components.

3.6 Direct Method

In the above methods, several mass components are assumed a priori, each of which had representative functional form, and therefore, the results depend on the assumed profiles. In order to avoid this dependence on the adopted functions, a direct method to obtain the mass distribution without assuming the functional form can be applied, where the mass distribution is calculated directly using the observed rotation velocity data. Since the rotation curve is restricted to give force in the galactic plane, the method cannot give the three-dimensional information. Hence, a "true" mass profile in a real disk galaxy is assumed to lie between two extreme profiles, which are either spherical or axisymmetric flat-disk.

3.6.1 Spherical Mass Distribution

The mass $M(R)$ of a spherical body inside radius R is given by

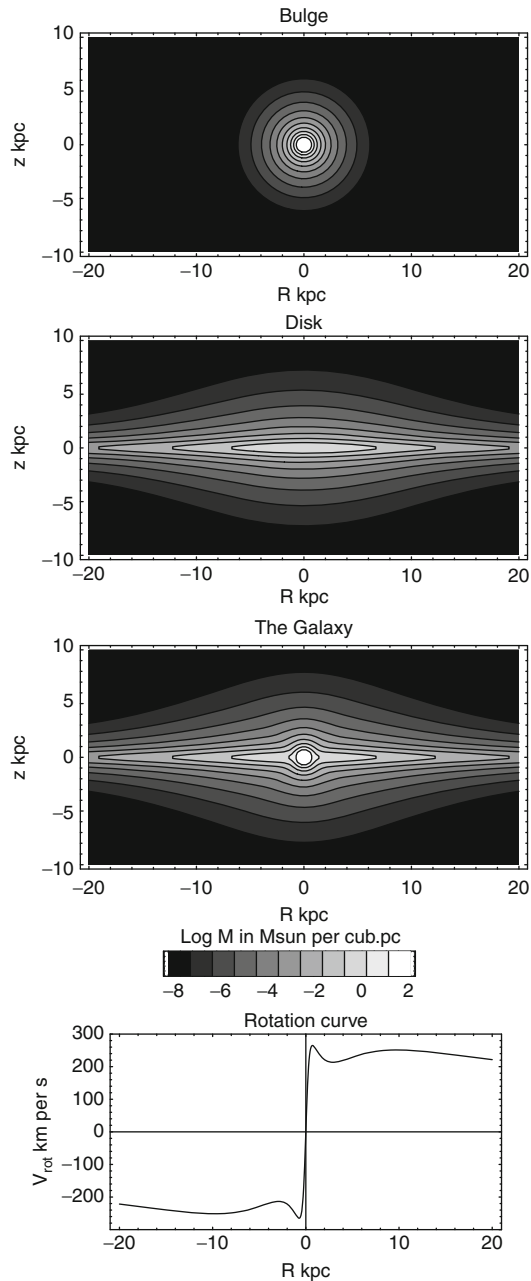
$$M(R) = \frac{RV(R)^2}{G}, \quad (19.43)$$

where $V(R)$ is the rotation velocity at r . Then the SMD $\Sigma_S(R)$ at R is calculated by,

$$\Sigma_S(R) = 2 \int_0^{\infty} \rho(r) dz = \frac{1}{2\pi} \int_R^{\infty} \frac{1}{r\sqrt{r^2 - R^2}} \frac{dM(r)}{dr} dr. \quad (19.44)$$

The volume mass density $\rho(R)$ is given by

$$\rho(R) = \frac{1}{4\pi r^2} \frac{dM(r)}{dr}. \quad (19.45)$$



■ Fig. 19-14

Meridional distributions of the volume density in the bulge, disk, and their superposition to represent the Galaxy calculated for the Miyamoto–Nagai potential (1975) with parameters in [Table 19-4](#). Contours and gray scales are logarithmic. The corresponding rotation curve is shown at *bottom*

■ Table 19-4

Parameters determining the Miyamoto–Nagai potential of the galaxy

Component	$M_i (M_\odot)$	a_i (kpc)	b_i (kpc)
Bulge	2.05×10^{10}	0.0	0.495
Disk	2.547×10^{11}	7.258	0.520

For a given rotation curve $V(R)$, (● 19.44) can be computed numerically. In a galaxy, this gives a good approximation for the central region where the spheroidal component dominates. On the other hand, the equation gives underestimated mass density near the outer edge at $R \sim R_{\max}$ because of the edge effect due to the finite radius of data points. The edge effect is negligible in the usual disk regions.

3.6.2 Flat-Disk Mass Distribution

The surface mass density (SMD) for a thin disk, $\Sigma_D(R)$, can be obtained by solving Poisson’s equation on the assumption that the mass is distributed in a flat disk with negligible thickness. It is given by

$$\Sigma_D(R) = \frac{1}{\pi^2 G} \left[\frac{1}{R} \int_0^R \left(\frac{dV^2}{dr} \right)_x K \left(\frac{x}{R} \right) dx + \int_R^\infty \left(\frac{dV^2}{dr} \right)_x K \left(\frac{R}{x} \right) \frac{dx}{x} \right], \quad (19.46)$$

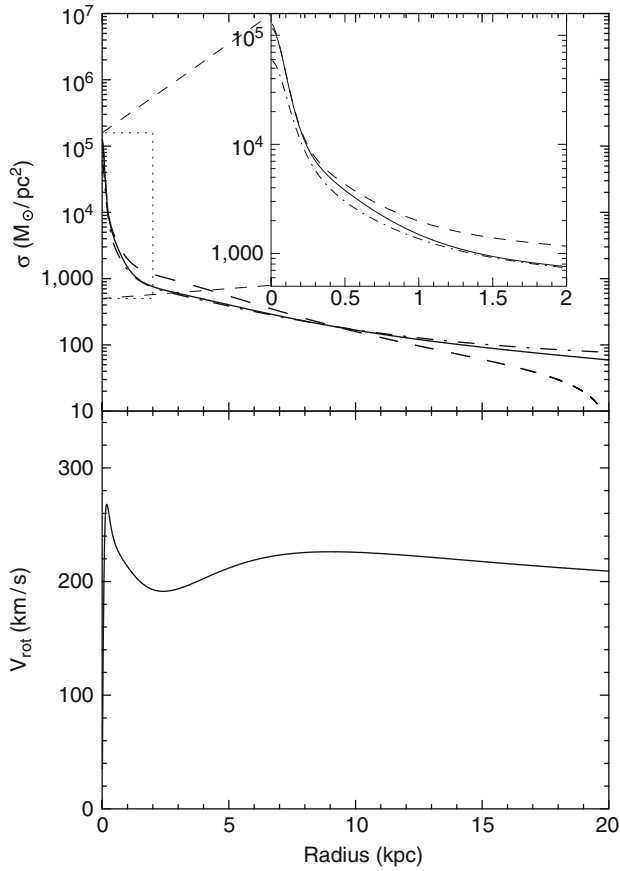
where K is the complete elliptic integral and becomes very large when $x \simeq R$ (Binney and Tremaine 1987). For the calculation, it must be taken into account that (● 19.46) is subject to the boundary condition, $V(0) = V(\infty) = 0$. Also it is assumed that $V(0) = 0$ at the center. Since a central black hole of a mass on the order of 10^6 – $10^7 M_\odot$ dominates the RC only within a few parsecs, it does not influence the galactic scale SMD profile.

When calculating the first term on the right-hand side of (● 19.46) for the central region, it happens that there exist only a few data points, where the reliability of the calculated $V(R)$ is lower than the outer region. In addition, the upper limit of the integration of the second term is R_{\max} instead of infinity. Since the rotation curves are nearly flat or declining outward from $R = R_{\max}$, the second term becomes negative. Thus, the values are usually slightly overestimated for $\Sigma_D(R)$ at $R \simeq R_{\max}$.

3.6.3 Verification Using the Miyamoto–Nagai Potential

It is interesting to examine how the results for the spherical and flat-disk assumptions, as computed from (● 19.44) and (● 19.46), respectively, differ from each other as well as from the “true” SMD for the Miyamoto–Nagai (1975) potential. Given a set of parameters, the “true” SMD as well as the rotation curve can be obtained. Using this rotation curve, the SMDs both for a spherical and flat-disk are calculated using the methods as described in the previous sections.

● Figure 19-15 shows the “true” SMD and computed SMDs from the rotation curve for spherical and flat-disk assumptions calculated. Here, the rotation curve only up to $R = 20$ kpc



■ Fig. 19-15

Comparison of the mass deconvolution (*top panel*) using spherical and flat-disk models applied to the analytical rotation curve for a Miyamoto–Nagai potential (*bottom panel*) (Takamiya and Sofue 2000)

is used. The figure demonstrates that the spherical case well reproduces the true SMD for the inner region. This is reasonable, because the spherical component is dominant within the bulge. On the other hand, the flat-disk case better reproduces the true SMD in the disk region, which is also reasonable. Near the outer edge, the flat-disk case better traces the true SMD, while the spherical case is affected by the edge effect significantly.

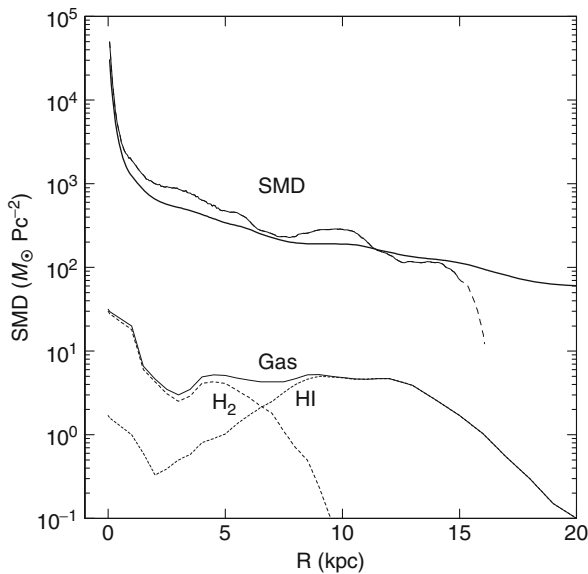
It must be stressed that the results from the two extreme assumptions, spherical and flat-disk, differ at most by a factor of 2, and do not differ by more than a factor of 1.5 from the true SMD in most regions except for the edge. It may be thus safely assumed that the true SMD is in between the two extreme cases. It should be remembered that the SMD is better represented by a spherical case for the inner region, while a flat-disk case is better for the disk and outer part.

3.7 Direct Mass Distribution in the Galaxy

► *Figure 19-16* shows the obtained direct SMD distribution in the Galaxy. There is remarkable similarity between this result and the total SMD obtained by deconvolution of the rotation curve into the bulge, disk, and halo components, as indicated by the upper dashed line in ► *Fig. 19-13*.

The mass is strongly concentrated toward the nucleus, and the bulge component dominates in the central region. The calculated SMD reaches a value as high as $\sim 10^5 M_\odot \text{pc}^{-2}$ at radius of a few tens parsecs. Higher density concentration has been observed from high-resolution infrared photometry and spectroscopy, indicating SMD as high as $\sim 10^6 M_\odot \text{pc}^{-3}$ within a few parsecs (Genzel et al. 1996). These values may be compared with the central value of the bulge's SMD as fitted by the de Vaucouleurs profile, $\Sigma_{\text{bc}} = 6.8 \times 10^6 M_\odot \text{pc}^{-2}$, from (► 19.27) (► *Table 19-2*).

The galactic disk appears as the straight-line part at $R \sim 3$ to 8 kpc on this semi-logarithmic plot (R vs $\log \Sigma$), indicating the exponential character. Even in these radii, as well as in the solar vicinity at $R \sim 8$ kpc, the dynamical surface mass (not volume density) is dominated by the dark matter, because the SMD is the projection of huge extent of the dark halo. There is slight difference between SMDs from the deconvolution method and direct method: for example at $R \sim 8$ kpc, the SMD is $\sim 300 M_\odot \text{pc}^{-2}$ in ► *Fig. 19-13*, whereas it is ~ 200 – $250 M_\odot \text{pc}^{-2}$ in ► *Fig. 19-16*. This discrepancy is due to the difference caused by the limited



■ Fig. 19-16

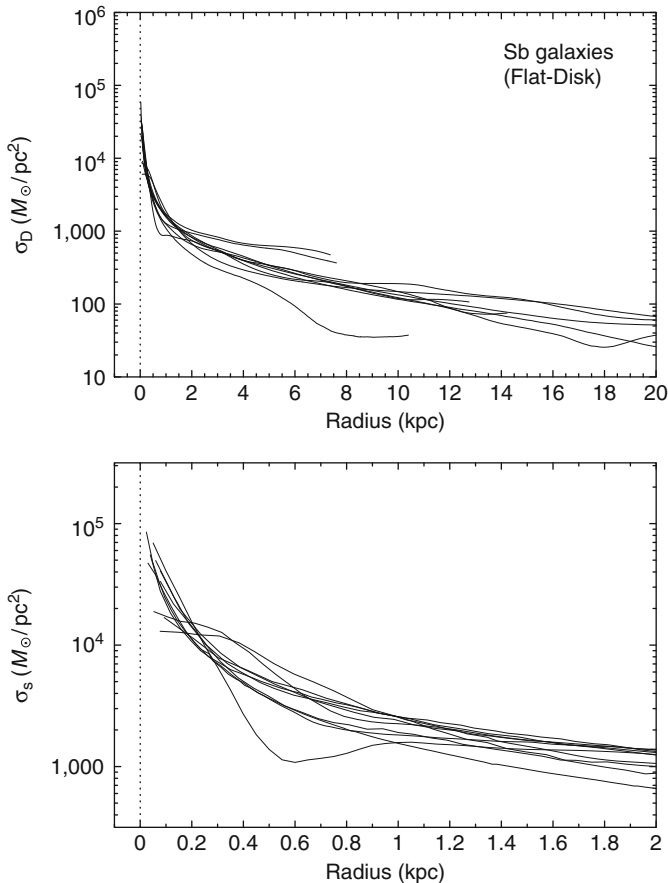
Radial distribution of the surface mass density (SMD) in the Galaxy directly calculated from the rotation curve on the disk assumption (*thick line*). The thin line is the result for spherical assumption which better represents the innermost bulge component. The outer truncation (*dashed part*) is due to edge effect. Lower thin line is SMD of interstellar gas made by annulus-averaging the face-on projected distribution in ► *Fig. 19-19* (Nakanishi and Sofue 2003, 2006). HI and H₂ gas SMDs are also shown separately by *dotted lines*

areas of integration from the finite data for the infinitely extended dark halo as well as due to difference in the adopted methods.

The outer disk, as indicated by the flat-disk model (full line in [Fig. 19-16](#)), is followed by an outskirts with a more slowly declining density profile, gradually detaching from the straight-line part of the disk. This outskirts indicates the dark halo, which extends to much larger radii, as is discussed in the next section.

3.8 Direct Mass Distributions in Spiral Galaxies

The mass distributions in spiral galaxies can be also directly computed from observed rotation curves in the same way as for the Milky Way. [Figure 19-17](#) shows the SMD distributions of Sb galaxies obtained for the rotation curves in [Fig. 19-10](#) using [\(19.46\)](#). The central regions are enlarged in the bottom panel, where [\(19.44\)](#) is adopted for a spherical model, which better



■ Fig. 19-17

(Top): Mass distributions in Sb galaxies using flat-disk. (Bottom): Same but for the central regions using spherical model (Takamiya and Sofue 2000)

represents the inner spheroidal components. The calculated SMD profiles for the Sb galaxies are similar to that of the Milky Way. It is also known that the profiles for Sa to Sc galaxies are very similar to each other, except for the absolute values (Takamiya and Sofue 2000). It is stressed that the dynamical structure represented by the density profile is very similar to each other among spiral galaxies. The SMD profiles have a universal characteristics as shown in these figures: high central concentration, exponential disk (straight line on the semilogarithmic plot), and outskirts due to the dark halo.

3.9 Distribution of Interstellar Gas in the Galaxy

The rotation curve of the Milky Way Galaxy is useful not only for deriving the mass distribution, but also for mapping the interstellar gas. Given a rotation curve $V(R)$, radial velocity v_r of any object near the galactic plane at $b \sim 0^\circ$ is uniquely calculated for its distance and longitude (l, r). Inversely, given the radial velocity and galactic longitude (v_r, l) of an object, its distance from the Sun, and therefore, its position in the galactic disk is determined. Thus, the distribution of objects and gases can be obtained by measuring radial velocities from spectroscopic observations such as of recombination lines (HII regions), $\lambda 21$ -cm line emission (HI clouds and diffuse gas), molecular lines (CO lines), and/or maser lines (e.g., SiO).

The radial velocity of an object in the galactic plane is calculated by

$$v_r = R_0(\omega - \omega_0)\sin l = \left\{ \frac{R_0}{R} V(R) - V_0 \right\} \sin l. \quad (19.47)$$

Using this equation, a radial-velocity diagram is obtained for the Galaxy as shown in [Fig. 19-18](#).

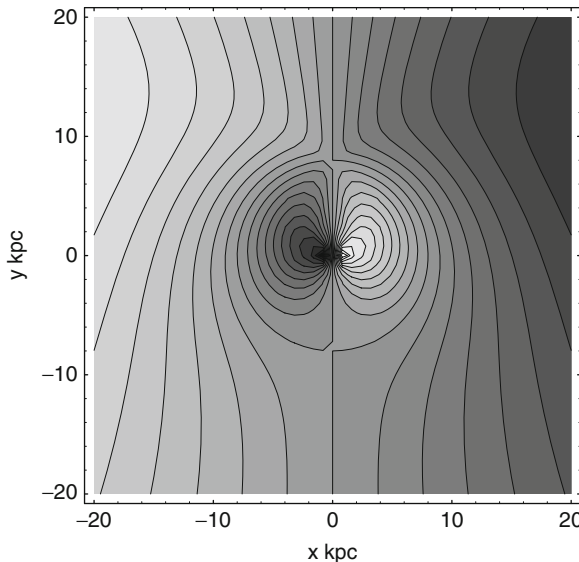


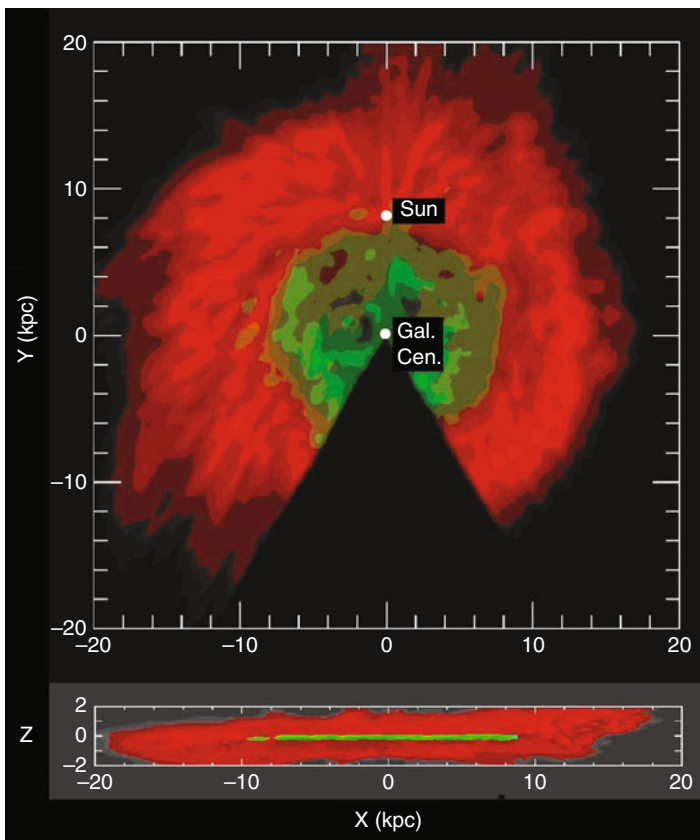
Fig. 19-18
Radial-velocity field in the galactic plane

The radial-velocity field is used to determine the position (distance r) of an object in the Galaxy for which the coordinates and radial velocity, $(l, b; v_r)$, are known. The volume density of interstellar gas at the position is obtained from the line intensity. This procedure is called the velocity-to-space transformation (VST) and is useful to produce a face-on map of the density distribution. Positions of objects outside the solar circle are uniquely determined by this method. However, inside the solar circle the solution for the distance is twofold, appearing either at near or at far side of the tangent point. For solving this problem, additional information such as apparent diameters of clouds and thickness of the HI and molecular disks are required.

With the VST method, the HI and H₂ gas maps can be obtained by the following procedure. The column density of gas is related to the line intensities as

$$N_{\text{HI,H}_2} [\text{H, H}_2 \text{ cm}^{-2}] = C_{\text{HI,H}_2} \int T_{\text{HI,CO}}(v) dv [\text{K km s}^{-1}], \quad (19.48)$$

where T_{HI} and T_{CO} are the brightness temperatures of the HI and CO lines, and $C_{\text{HI}} = 1.82 \times 10^{18}$ [H atoms cm^{-2}] and $C_{\text{H}_2} \sim 2 \times 10^{20} f(R, Z)$ [H₂ molecules cm^{-2}] are the conversion factors from the line intensities to column densities of HI and H₂ gases, respectively. Here, $f(R, Z)$ is



■ Fig. 19-19

Distribution of surface density of interstellar gas in the Galaxy obtained from spectral data of the HI 21-cm and CO-2.6-mm line emissions on an assumption of circular rotation of the gas (Nakanishi and Sofue 2003, 2006)

a correction factor ranging from 0.1 to 2 for the galactocentric distance R and metallicity Z in the galaxy with the solar-vicinity value of unity (Arimoto et al. 1996). The local density of the gas at the position that corresponds to the radial velocity v is observed is obtained by

$$n_{\text{HI,H}_2} = \frac{dN_{\text{HI,H}_2}}{dr} = \frac{dN_{\text{HI,H}_2}}{dv} \frac{dv}{dr} = C_{\text{HI,H}_2} T_{\text{HI,CO}}(v) dv/dr. \quad (19.49)$$

Combining the LV diagrams as shown in [Fig. 19-2](#) and the velocity field, a “face-on” distribution is obtained of the density of interstellar gas in the Galaxy. [Figure 19-19](#) shows a face-on view of the Galactic disk as seen in the HI and molecular line emissions (Nakanishi and Sofue 2003, 2006). [Figure 19-16](#) shows azimuthally averaged radial profiles of the interstellar gas density. The gaseous mass density is far less compared to the dynamical mass densities by stars and dark matter by an order of magnitude. The interstellar gas density at $R \sim 8$ kpc is $\sim 5.0 M_{\odot} \text{pc}^{-2}$, which shares only several percents of disk mass density of $\sim 87.5 M_{\odot} \text{pc}^{-2}$ ([Figs. 19-13](#) and [19-16](#), [Table 19-3](#)).

4 Dark Halo

4.1 Dark Halo in the Milky Way

The mass of the Galaxy inside the solar circle is $\approx 10^{11} M_{\odot}$. The mass interior to the distance of the Large Magellanic Cloud at 50 kpc may grow to $6 \times 10^{11} M_{\odot}$ (Wilkinson and Evans 1999). Interior to 200 kpc, the mass would be at least $2 \times 10^{12} M_{\odot}$ (Peebles 1995). The Milky Way and M31 are the major members of the Local Group, around which many satellites and dwarf galaxies are orbiting (Sawa and Fujimoto 2005). Masses and extents of dark halos around these two giant galaxies are crucial for understanding the dynamics and structure of the Local Group. Outer rotation curve of the Galaxy is a key to put constraints on the dark halo structure. Although the outer rotation curve is reasonably fitted by an isothermal model, the halo model might not be unique, and is difficult to clearly discriminate the isothermal model (Begeman et al. 1991) from NFW (Navarro et al. 1996) and Burkert models (Burkert 1995), because of the large scatter of data as well as for the limited radius within which the rotation curve is observed.

Kinematics of satellites and member galaxies in the Local Group is a crew to estimate the dark halo, which is considered to be extended far outside the galactic disk and in the intracluster space (Kahn and Woltjer 1959; Li and White 2008; van der Marel and Guhathakurta 2008). Sawa and Fujimoto (2005) have computed the past probable orbits of the major members of the Local Group under the condition that their positions and radial velocities are satisfied at the present time. In this section, the behaviors of rotation curves calculated for different dark halo models are examined, and are compared with a “pseudo rotation curve” of the Local Group, which combines the rotation curve of the Galaxy and radial velocities of the members of the Local Group.

4.1.1 Isothermal Halo Model

The simplest interpretation of the flat rotation curve observed in the outer Galaxy, and outer rotation curves in many spiral galaxies is to adopt the semi-isothermal spherical distribution for the dark halo (Begeman et al. 1991; Kent 1986). In the isothermal model, the density profile

is written as

$$\rho_{\text{iso}}(R) = \frac{\rho_{\text{iso}}^0}{1 + (R/h)^2}, \quad (19.50)$$

where ρ_{hc} and $h = R_h$ are constants giving the central mass density and scale radius of the halo, respectively. This profile gives finite mass density at the center, but yields a flat rotation curve at large radius. The circular velocity is given by

$$V_h(R) = V_\infty \sqrt{1 - \left(\frac{h}{R}\right) \tan^{-1}\left(\frac{R}{h}\right)}, \quad (19.51)$$

where V_∞ is a constant giving the flat rotation velocity at infinity. The constants are related to each other as

$$V_\infty = \sqrt{4\pi G \rho_{\text{iso}}^0 h^2}, \quad (19.52)$$

or the central density is written as

$$\rho_{\text{iso}}^0 = 0.740 \left(\frac{V_\infty}{200 \text{ km s}^{-1}} \right) \left(\frac{h}{1 \text{ kpc}} \right)^{-2} M_\odot \text{ pc}^{-3}. \quad (19.53)$$

The enclosed mass within radius R is given by

$$M(R) = 4\pi \int_0^R \rho_i(r) r^2 dr \quad (19.54)$$

with $i = \text{iso}$ (NFW or Burkert for the other two models as discussed below).

At small radius, $R \ll h$, the density becomes nearly constant equal to ρ_{iso}^0 and the enclosed mass increases steeply as $M(R) \propto R^3$. At large radius of $R \gg h$ the density decreases as $\rho_{\text{iso}} \propto R^{-2}$ and the enclosed mass tends to increase linearly with radius as $M(R) \propto R$.

4.1.2 NFW and Burkert Profiles

Based on numerical simulations of the formation of galaxies in the cold-dark matter scenario in the expanding universe, several model profiles have been found to fit better the calculated results. The most well-known model is the NFW model proposed by Navarro et al. (1996), and Burkert (1995) has modified this model. The NFW and Burkert density profiles are written, respectively, as

$$\rho_{\text{NFW}}(R) = \frac{\rho_{\text{NFW}}^0}{(R/h)[1 + (R/h)]^2}, \quad (19.55)$$

and

$$\rho_{\text{Bur}}(R) = \frac{\rho_{\text{Bur}}^0}{[1 + (R/h)][1 + (R/h)^2]}. \quad (19.56)$$

The circular rotation velocity is calculated by

$$V_h(R) = \sqrt{\frac{GM_h(R)}{R}}, \quad (19.57)$$

where M_h is the enclosed mass within h as calculated by (19.54).

At small radius with $R \ll h$, the NFW density profile behaves as $\rho_{\text{NFW}} \propto 1/R$, yielding an infinitely increasing density toward the center, and the enclosed mass behaves as $M(R) \propto R^2$. On the other hand, the Burkert profile tends to constant density ρ_{Bur}^0 , yielding steeply increasing

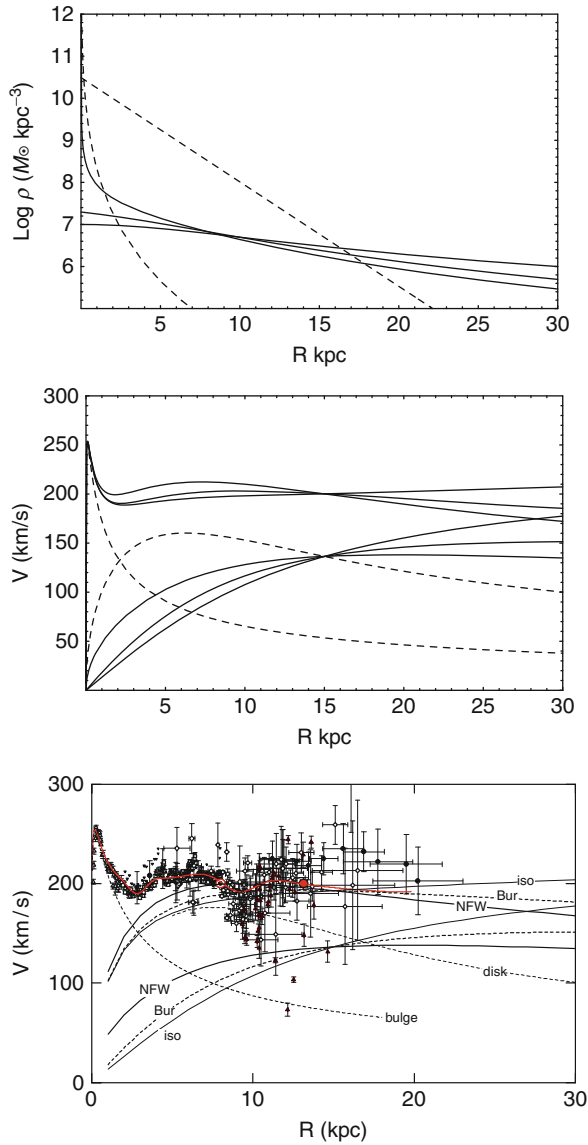
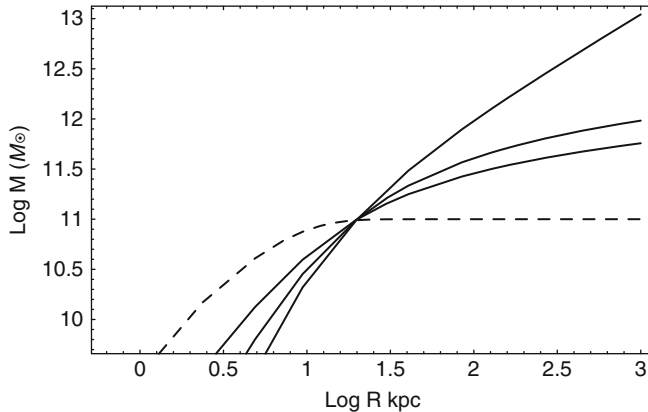


Fig. 19-20

(Top) Schematic volume density profiles for dark halo models – isothermal, Burkert, and NFW models (full lines from top to bottom at $R = 30$ kpc) – compared with the disk and bulge (dashed lines). (Middle) Rotation curves. (Bottom) Rotation curves compared with observations

enclosed mass as $M(R) \propto R^3$, similarly to the isothermal profile. At large radius with $R \gg h$, both the NFW and Burkert profiles have densities $\rho_{\text{NFW, Bur}} \propto R^{-3}$, and they yield milder logarithmic increase of mass as $M(R) \propto \ln R$ (Fig. 19-21).

The scale radius of a dark halo model is usually supposed to be between 3.5 and 10 kpc. Here, a value of $h = 10$ kpc is adopted commonly for the three models. Figure 19-20 shows



■ Fig. 19-21

Enclosed masses of dark halos up to 1 Mpc (isothermal, Burkert, and NFW from top to bottom at 1 Mpc), compared with the disk mass (*dashed line*) that approximately manifests the galactic luminous mass

density distributions for the three different models of the dark halo as well those for the disk and bulge. Corresponding rotation curves are then compared with the observations. Figure 19-20 also shows the masses enclosed within a sphere of radius R . In these figures, the total masses of the disk and halo are taken to be the same at $R = 15$ kpc, so that the rotation velocity is nearly flat at $R = 10$ – 20 kpc and fit the observations. Figure 19-21 calculates the enclosed masses up to radius of 1 Mpc for the three models. This figure demonstrates that the isothermal model predicts rapidly increasing mass merging with the neighboring galaxies' halos, whereas the NFW and Burkert models predict a rather isolated system with mild (logarithmic) increase of enclosed mass for the Milky Way.

4.1.3 Dark Halo Contribution in the Inner Galaxy

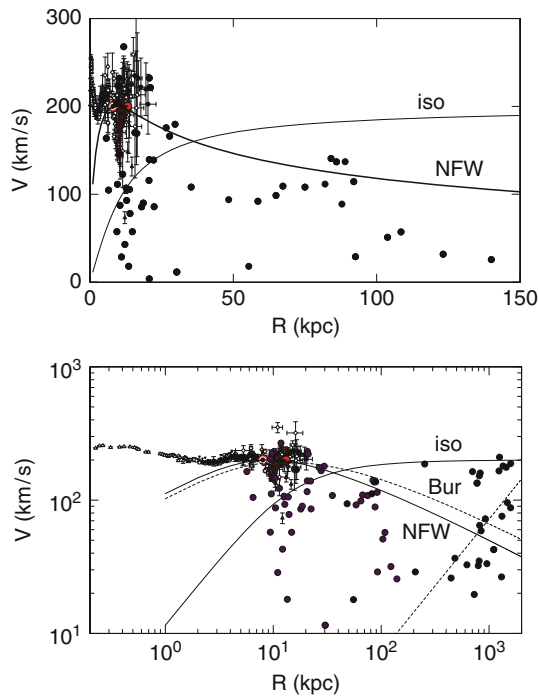
In the inner Galaxy at $R < \sim 10$ kpc, the rotation velocity is predominantly determined by the bulge and disk contributions. Among the three dark halo profiles, only the NFW model predicts a nuclear cusp in which the dark matter density increases toward the nucleus inversely proportional to the radius. However, the bulge density is kept sufficiently large compared to the dark halo density in the Galactic Center. The Burkert and isothermal halo models indicate a mild density plateau at the center, which is also much smaller than the disk and bulge densities.

Accordingly, in any models, the contribution from the dark halo to the rotation velocity is negligible in the inner Galaxy. Thus, it is hard to discriminate the dark halo models by comparing with the observations in the inner Galaxy, although the data are most accurately obtained there. The rotation velocity becomes sensitive to the dark halo models only beyond ~ 20 kpc.

4.2 Pseudo Rotation Curve of the Local Group

In order to discuss the mass distribution in the dark halo, kinematics of objects surrounding the Galactic disk is essential (e.g., Kahn and Woltjer 1959; Sawa and Fujimoto 2005). A method for deriving dark halo density profiles by using radial velocities of the Local Group of galaxies as an extension of the rotation curve of the Galaxy is described below.

In [Fig. 19-22](#), absolute values of galactocentric radial velocities $V_r = |V_{GC}|$ of outer globular clusters and member galaxies of the Local Group are plotted against their galactocentric distances R_{GC} (Sofue 2009), and the literature cited therein for the data. In the figure, calculated rotation curves for the NFW model is shown by the thick line, and individual components are shown by dashed lines. The isothermal halo model is shown by the upper thin line, which is horizontal at large radius. The upper boundaries of the plot are well fitted by the rotation curve for the NFW model. [Figure 19-22](#) shows the same for entire region of the Local Group up to 1 Mpc. The galactocentric distance R_{GC} was calculated from galactic coordinates



■ Fig. 19-22

Pseudo rotation curves within 150 kpc (*top*) and 1.5 Mpc in logarithmic distance scale (*bottom*). Data beyond the galactic disk are absolute values of galactocentric radial velocities of globular clusters, satellite galaxies, and Local Group members. The big circle at 770 kpc denotes M31. The thick line shows the rotation curve for the NFW halo model, which traces well the upper boundary of the plot at $R \leq 150$ kpc. The asymptotic horizontal thin line represents the isothermal model. Hubble flow for $H_0 = 72 \text{ km s}^{-1} \text{ Mpc}^{-1}$ is shown by a dashed line near the right-bottom corner

and heliocentric distance, and V_r was calculated from observed heliocentric radial velocity by correcting for the solar rotation of $V_0 = 200 \text{ km s}^{-1}$ at $R = R_0$.

In these figures, the calculated rotation curves for the three dark halo models are superposed. The dark halo models are so chosen that they are smoothly connected to the inner composite rotation curve for the best-fit bulge and disk. The scale radius of the dark halo is taken to be $h = 10 \text{ kpc}$. The total mass of the dark halo within the critical radius $R_c = 15 \text{ kpc}$ in each model is taken to be equal to the disk mass in the same radius.

As equations (19.55) and (19.56) indicate, the NFW and Burkert profiles are similar to each other at radii sufficiently greater than the scale radius at $R \gg h \text{ kpc}$. Accordingly, the rotation curves corresponding to these two models are similar at $R > 30 \text{ kpc}$.

If the area at $R < \sim 150 \text{ kpc}$ in Fig. 19-22 are carefully inspected, the NFW traces the upper envelope of plot better than isothermal profile. On the other hand, the isothermal model can better trace the upper boundary of the entire observations in the Local Group up to $\sim 1.5 \text{ Mpc}$. M31, one of the major two massive galaxies in the Local Group, lies on the isothermal curve. These facts imply that the mass concentration around the Galaxy within $\sim 150 \text{ kpc}$ is isolated from the larger-scale mass distribution controlling the whole Local Group members. Hence, if it is gravitationally bound to the Local Group, there exists a much larger amount of dark mass than the dark halo of the Galaxy, filling the entire Local Group.

4.3 Mass of the Galaxy Embedded in the Dark Halo

The upper boundary of the pseudo rotation curve within $\sim 150 \text{ kpc}$ radius is well fitted by the NFW and Burkert models, as shown in Fig. 19-22. In these models, the total mass of the Galaxy involved within a radius 770 kpc , the distance to M31, is $8.7 \times 10^{11} M_\odot$, and the mass within 385 kpc , a half way to M31, is $4.4 \times 10^{11} M_\odot$, respectively. These values may be adjustable within a range less than $10^{12} M_\odot$ by tuning the parameters h and R_c , so that the flat galactic rotation at $R \sim 10\text{--}20 \text{ kpc}$ can be approximated.

In either case, the Galaxy's mass is by an order of magnitude smaller than the total mass required for gravitational binding of the Local Group. If the two galaxies, M31 and the Galaxy, are assumed to be gravitationally bound to the Local Group, the total mass is estimated to be on the order of $M_{\text{tot}} \sim V^2 R/G \sim 4.8 \times 10^{12} M_\odot$, as estimated from the mutual velocity of the two galaxies of $V \sim \sqrt{3} V_{\text{GC}}$ with $V_{\text{GC}} = -134 \text{ km s}^{-1}$ and the radius of the orbit is $R = 385 \text{ kpc}$. Here, the factor $\sqrt{3}$ was adopted as a correction for the freedom of motion. Thus, the Local Group may be considered to contain a dark matter core of mass comparable to this binding mass, $\sim 4.8 \times 10^{12} M_\odot$. The mean mass density required to stabilize the Local Group is, thus, estimated to be $\rho_{\text{LG}} \sim 2 \times 10^{-5} M_\odot \text{pc}^{-3}$.

Now, the boundary of the Galaxy is defined as the radius, at which the density of the galactic dark halo becomes equal to the dark matter density of the Local Group. The thus defined radius is $R_G \sim 100 \text{ kpc}$ for the NFW model. The Galaxy's mass within this boundary is $M_{\text{G:NFW}} = 3 \times 10^{11} M_\odot$ (Fig. 19-20).

The high-velocity ends of the pseudo rotation curve for the entire Local Group up to $\sim 1.5 \text{ Mpc}$ in Fig. 19-22 are well represented by an isothermal dark halo model. The estimated total mass is $M \sim 4.8 \times 10^{12} M_\odot$ for the terminal flat velocity of $V = 134\sqrt{3} \sim 230 \text{ km s}^{-1}$. This may be compared with the estimates by Li and White (2008) ($5.3 \times 10^{12} M_\odot$) and van der Marel and Guhathakurta (2008) ($5.6 \times 10^{12} M_\odot$).

The mean dark matter density of the Universe is on the order of

$$\rho_{\text{uni}} \sim \frac{(H_0 R)^2 R}{G} \frac{1}{\frac{4\pi}{3} R^3} = \frac{3H_0^2}{4\pi G}, \quad (19.58)$$

which is approximately $\sim 2 \times 10^{-29} \text{ g cm}^{-3} \sim 1.2 \times 10^{-6} M_\odot \text{ pc}^{-3}$ for $H_0 = 72 \text{ km s}^{-1} \text{ kpc}^{-1}$. The total mass of this “uniform” dark matter is $\sim 1.2 \times 10^{11} M_\odot$ inside a sphere of radius 385 kpc, and $\sim 1 \times 10^{12} M_\odot$ in 770 kpc, which is small enough compared to the dynamical mass of the Local Group.

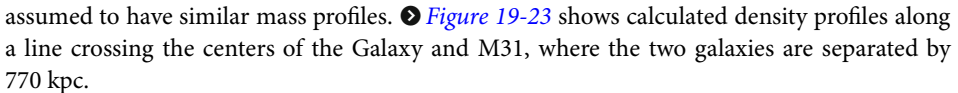
The ratio of the baryonic mass (luminous mass) of a galaxy to the dark matter mass gives important information for the formation scenario of galaxies. The upper limit to the baryonic mass of the Galaxy is approximately represented by the disk plus bulge mass, which was estimated to be $M_{\text{baryon}} \simeq 0.83 \times 10^{11} M_\odot$ by fitting of the inner rotation curve using de Vaucouleurs bulge and exponential disk. This yields an upper limit to the baryon-to-dark matter mass ratio of the Galaxy within radius R to be $\Gamma(R) \simeq 0.38$ for $R = 100 \text{ kpc}$, and $\Gamma \simeq 0.23$ for $R = 385 \text{ kpc}$, where Γ is defined by

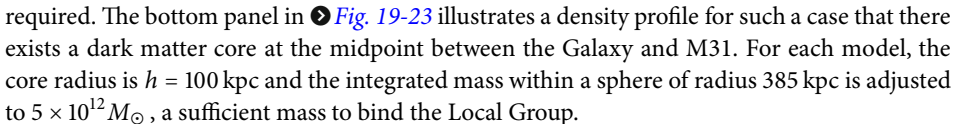
$$\Gamma(R) = \frac{M_{\text{baryon}}(R)}{M_{\text{dark matter}}(R)} = \frac{M_{\text{baryon}}(R)}{\{M_{\text{total}}(R) - M_{\text{baryon}}(R)\}}. \quad (19.59)$$

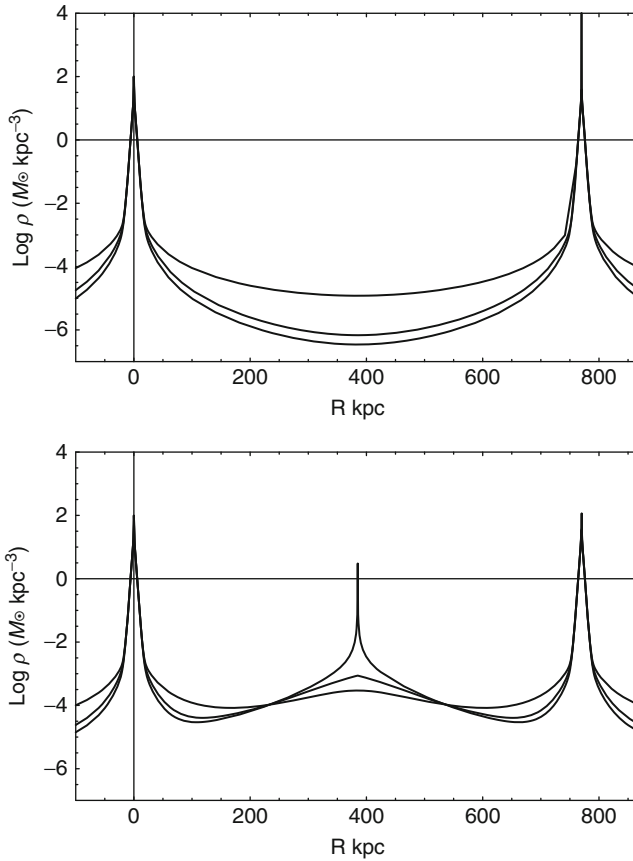
These values are on the same order of the cosmological baryon-to-dark matter ratio of 0.20 (=4.6%/23%) from WMAP (Spergel et al. 2003). Thus, the baryon-to-dark matter ratio within the Galaxy’s boundary is close to the cosmological value.

The baryonic mass of galaxies in the Local Group may be further compared with the total Local Group mass. The total luminous mass is represented by those of the Galaxy and M31, which is about twice the Galaxy’s baryonic mass, because the contribution from the other small and dwarf galaxies are negligible. Then, the upper limit to the baryon-to-dark matter mass ratio in the whole Local Group is estimated to be $\Gamma \sim 0.036$, which is only 0.2 times the cosmic value. It is expected that approximately 80% of baryonic matter of the Local Group exists in the intergalactic space without being captured to the present galaxies.

4.4 The Galaxy, M31, and Local Group

HI observations by showed that M31’s rotation curve is flat till a radius 35 kpc. They estimate the mass of M31 within 35 kpc to be $3.4 \times 10^{11} M_\odot$. This is comparable to the total mass of the Milky Way Galaxy within 35 kpc of $3.2 \times 10^{11} M_\odot$. The Galaxy and M31 may be assumed to have similar mass profiles.  **Figure 19-23** shows calculated density profiles along a line crossing the centers of the Galaxy and M31, where the two galaxies are separated by 770 kpc.

If the halo profile is either Burkert or NFW, which is favored from the observations inside $R \sim 150 \text{ kpc}$, the two galaxies are not massive enough to bind each other. In order to stabilize the Local Group, an extended massive component of intracluster dark matter of several $\times 10^{12} M_\odot$ is required. The bottom panel in  **Fig. 19-23** illustrates a density profile for such a case that there exists a dark matter core at the midpoint between the Galaxy and M31. For each model, the core radius is $h = 100 \text{ kpc}$ and the integrated mass within a sphere of radius 385 kpc is adjusted to $5 \times 10^{12} M_\odot$, a sufficient mass to bind the Local Group.



■ Fig. 19-23

(Top) Density profiles between the Milky Way and M31 for three dark halo models – isothermal, Burkert, and NFW – from top to bottom. The isothermal halo (upper curve) gravitationally binds the two galaxies, while the other two potentials cannot bind the two galaxies. (Bottom) If the Galaxy and M31 are located on both sides of a dark matter core of the Local Group, the dark matter can bind the two galaxies

Less massive galaxies are floating in the dark matter ocean between the two big continents, M31 and the Galaxy. It is known that rotation curves of dwarfs and latest type galaxies increase monotonically toward their edges, as is typically observed for M33 (Corbelli and Salucci 2000). M33 is speculated to border M31 by their dark halos, and the rotation velocity increases until it reaches the velocity dispersion in the Local Group of $\sim 200 \text{ km s}^{-1}$.

A question is encountered about “temperatures” of dark matter. Inside small galaxies, rotation velocities are as small as $\sim 100 \text{ km s}^{-1}$, much less than those in giant spirals and intracluster space, and therefore, the potential is shallow. Accordingly, the dark matter is cooler, so that it is gravitationally bound to the shallow potential. On the other hand, the halo continuously merges with halos of bigger galaxies and/or intracluster matter with higher velocities

of $\sim 200 \text{ km s}^{-1}$. Thus, the temperature of dark matter increases from inside to outside of a dwarf galaxy. This temperature transfer of dark matter occurs also from inside the outskirts of the Galaxy at $R \sim 100\text{--}150 \text{ kpc}$, where the rotation velocity is $\sim 100 \text{ km s}^{-1}$ according to the well-fitted NFW model, to outside where the Local Group dark matter with velocity dispersion of $\sim 200 \text{ km s}^{-1}$ is dominant.

It is considered that there are three components of dark matter. First, the galactic dark matter having the NFW profile, which defines the mass distribution in a galaxy controlling the outer rotation curve; second, extended dark matter filling the entire Local Group having a velocity dispersion as high as $\sim 200 \text{ km s}^{-1}$, which gravitationally stabilize the Local Group; and third, uniform dark matter having much higher velocities originating from super galactic structures. The third component, however, does not significantly affect the structure and dynamics of the present Local Group. It is therefore speculated that at any place in the Galaxy, there are three different components of dark matter having different velocities, or different temperatures. They may behave almost independently from each other, but are interacting by their gravity.

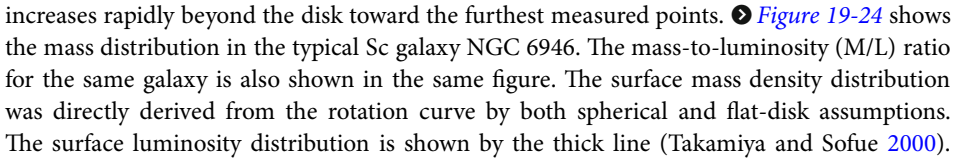
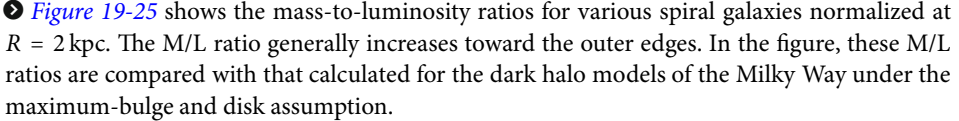
4.5 Dark Halos in Galaxies

By the 1970s, it had been recognized that rotation curves of spiral galaxies are flat to distances as large as $R \sim 30\text{--}50 \text{ kpc}$ from the nuclei by spectroscopy of HII regions (Rubin et al. 1982, 1985, 1997, 1999) and HI-line emission (Roberts and Rots 1973). The flat rotation curves suggested that the dynamical mass continued to rise to the last measured regions of the galaxies. Theoretically, dark matter halos were postulated to exist in any spiral galaxies in order for the disk to be stable from gravitational fragmentation (Ostriker and Peebles 1973). By analyzing motions of satellite and companion galaxies, Einasto et al. (1974) had shown that spiral galaxies are surrounded by massive dark halos.

Deeper and higher-resolution HI observations with synthesis telescopes reveal that for the majority of spiral galaxies, rotation curves remain flat beyond the optical disks (Bosma 1981a, b; Begeman 1989; van Albada et al. 1985). The largest HI disk has been known for Sc galaxy UGC 2885 with HI radius of 83 kpc for $H_0 = 72 \text{ km s}^{-1} \text{ Mpc}^{-1}$ where the rotation curve is still flat (Roelfsema and Allen 1985).

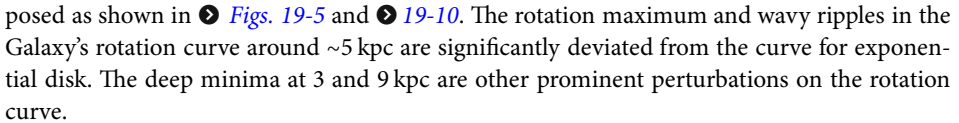
4.6 Mass-to-Luminosity Ratio

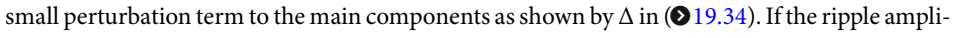
The best indicator of dark matter in a galaxy is the difference between the galaxy mass predicted by the luminosity and the mass predicted by the rotation velocities. This difference is usually indicated by the ratio of the dynamical mass to luminosity, which is called the mass-to-luminosity ratio (M/L). The M/L ratio is a clue to the distribution of visible and dark mass. Most investigations have assumed that the luminous part of a galaxy consists of bulge and disk, each with a constant mass-to-luminosity ratio. The “maximum-disk” assumption, that the disk component corresponding the rotation curve is dominated totally by baryonic mass, is often adopted to derive M/L ratios in the individual components (Kent 1986, 1992; Takamiya and Sofue 2000).

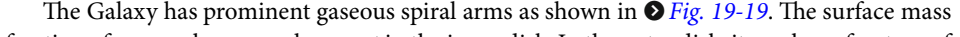
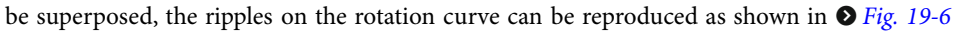
The distribution of M/L ratio in a galaxy is obtained by measuring the surface mass density (SMD) and surface luminosity profiles (Forbes 1992; Takamiya and Sofue 2000). The observations have indicated that the M/L ratio is highly variable within individual galaxies, and it increases rapidly beyond the disk toward the furthest measured points.  *Figure 19-24* shows the mass distribution in the typical Sc galaxy NGC 6946. The mass-to-luminosity (M/L) ratio for the same galaxy is also shown in the same figure. The surface mass density distribution was directly derived from the rotation curve by both spherical and flat-disk assumptions. The surface luminosity distribution is shown by the thick line (Takamiya and Sofue 2000).  *Figure 19-25* shows the mass-to-luminosity ratios for various spiral galaxies normalized at $R = 2$ kpc. The M/L ratio generally increases toward the outer edges. In the figure, these M/L ratios are compared with that calculated for the dark halo models of the Milky Way under the maximum-bulge and disk assumption.

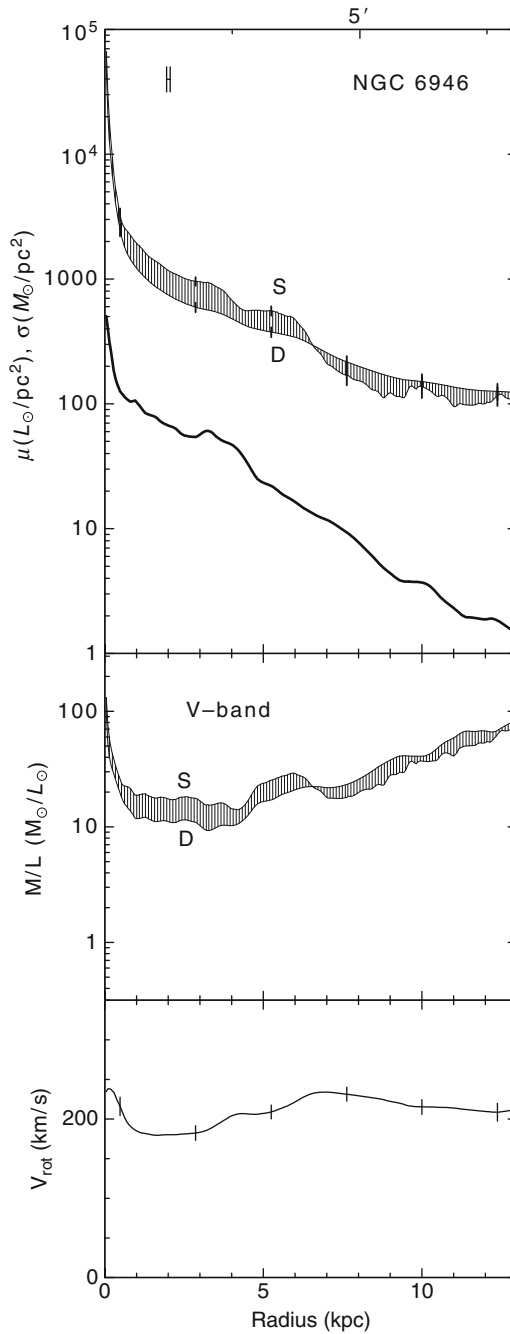
5 Smaller Mass Structures

5.1 Spiral Arms and Rotation Dips

Although rotation curves of galaxies are generally smooth and flat, wavy ripples are often superposed as shown in  *Figs. 19-5* and *19-10*. The rotation maximum and wavy ripples in the Galaxy's rotation curve around ~ 5 kpc are significantly deviated from the curve for exponential disk. The deep minima at 3 and 9 kpc are other prominent perturbations on the rotation curve.

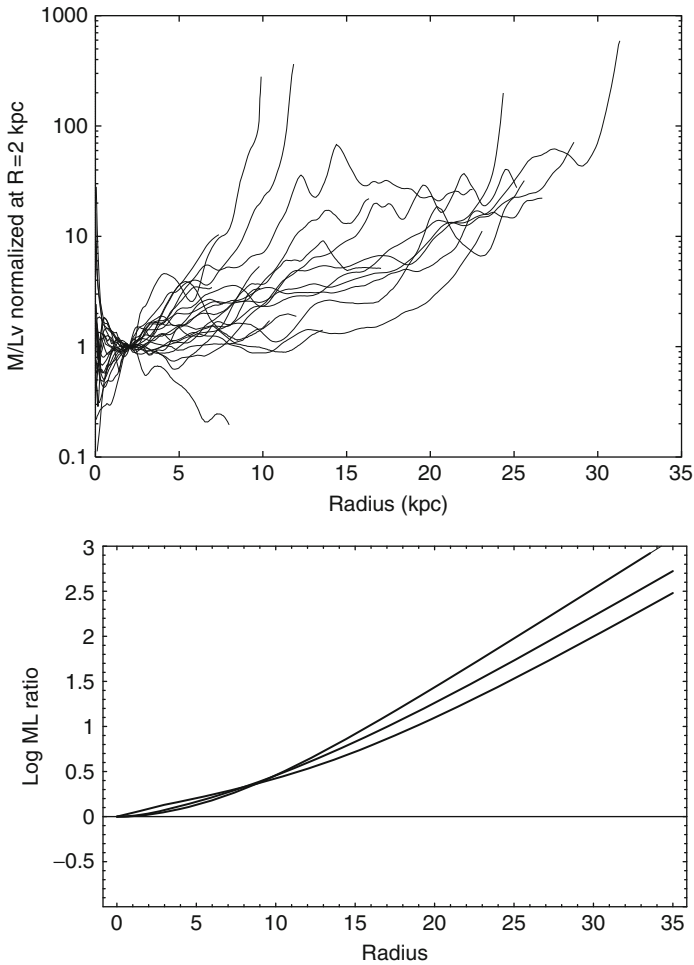
In addition to the smooth and axisymmetric dynamical components by the bulge, disk, and dark halo, which are the fundamental structure, the effects by smaller-scale structures such as spiral arms and bar should be considered. If the amplitude of the ripple is not large, they may be treated as a linear perturbation on the main structure, and can be represented by adding a small perturbation term to the main components as shown by Δ in  *19.34*. If the ripple amplitude is large and comparable to the rotation velocities, more sophisticated modeling is needed for nonlinear perturbation. There have been a number of numerical simulations to discuss the effect of a bar and spiral arms on the kinematics of interstellar gas and stars (e.g., Athanassoula 1992; Wada and Habe 1995). However, fitting to the observations by such nonlinear numerical simulation is not straightforward, and the results do not necessarily give unique solutions about the mass and shape of the perturbed structure. Here, only a brief discussion is given of qualitative properties noticed on the rotation curve, and is interpreted by simple analyses in a linear approximation.

The Galaxy has prominent gaseous spiral arms as shown in  *Fig. 19-19*. The surface mass fraction of gas reaches several percent in the inner disk. In the outer disk, it reaches a few tens of percent, but the mass of the dark halo much exceeds the disk. Hence, the contribution from the gaseous mass to the circular velocity is negligible in any region of the Galaxy. This means that the observed ripples of $\sim \pm 10$ km s⁻¹ on the rotation curve cannot be attributed to the gaseous arms alone. Some underlying more massive structures may be considered. If logarithmic spiral arms of density perturbation of a few tens of percent superposed on the disk are assumed to be superposed, the ripples on the rotation curve can be reproduced as shown in  *Fig. 19-6* (Sofue et al. 2009). In order to obtain a better fit to the observations including the 9 kpc dip,



■ Fig. 19-24

(Top) Surface mass density distribution of the spiral galaxy NGC 6946 calculated from the rotation curve by spherical and flat-disk assumptions. The surface luminosity distribution is shown by the thick line. (Middle): M/L ratio. (Bottom): Rotation curve (Takamiya and Sofue 2000)



■ Fig. 19-25

(Top) M/L ratios for spiral galaxies normalized at $R = 2$ kpc (Takamiya and Sofue 2000). (Bottom) M/L ratio approximately calculated for the dark halo models in the Galaxy. From upper to lower lines: isothermal, NFW, and Burkert models, respectively

a ring or a density wave of radius 11 kpc with amplitude as high as ~ 0.3 times the underlying disk density is required. It is possible that the dip is related to a density wave corresponding to the Perseus Arm.

5.2 Bar

The effect of a bar on the kinematics is significant in the inner galaxy (Athanasoula 1992), and observations have indeed revealed a bar in our Galaxy (Blitz et al. 1993; Weiland et al. 1994; Freudenreich 1998). Nakai (1992) noticed asymmetry in the distribution and kinematics of the CO-line emission at $360^\circ \leq l \leq 30^\circ$ with respect to the Galactic Center (► Fig. 19-2), and argued

that it may be attributed to noncircular motion by a bar. Similar bar-driven noncircular motions are also observed in molecular disks of many spiral galaxies (Kenney et al. 1992; Sakamoto et al. 1999).

The rotation dip near 3 kpc may be discussed in the scheme of nonlinear response of gas to the barred potential. According to near-infrared observations (Freudenreich 1998), the bar length is ~ 1.7 kpc and the tilt angle of 13° from the Sun-Galactic center line. If the bar's amplitude is taken to be on the order of about $\delta\text{bar} \simeq 20\%$ of the background smooth component, the 3 kpc dip in the observed rotation curve is well explained as due to noncircular motion by the bar's potential (🔍 Fig. 19-6). The excess mass by the bar may be estimated to be on the order of $M_{\text{bar}} \sim \delta M_{\text{disk+bulge}}(R \leq 2\text{kpc}) \sim 10^{10} M_\odot$.

In more inner region, observations show that the majority ($\sim 95\%$) of gas in the Galactic Center at $R \leq 300$ pc is rotating in a ring showing rigid-body features on the PV diagram (Bally et al. 1987; Oka et al. 1998; Sofue 1995). A few percent of the gas exhibits forbidden velocities, which may be either due to noncircular motion in an oval potential (Binney et al. 1991) or to expanding motion. The fact that the majority of the gas is regularly rotating indicates that the gas is on circular orbits.

5.3 Massive Central Component and Black Hole

Central rotation curves and mass distributions have been produced for a number of galaxies by a systematic compilation of PV diagrams in the CO and H α lines (Bertola et al. 1998; Sofue 1996, 2000). When they are observed with sufficiently high resolutions, innermost velocities of galaxies start already from high values at the nucleus, indicating the existence of high-density cores in their nuclei. High central density is not a characteristic only for massive galaxies. The nearby Sc galaxy M33, having a small bulge, exhibits central velocities at about $V=100$ km s $^{-1}$ (Rubin and Graham 1987) which do not decrease to zero at the origin. Thus, in spiral galaxies, the contribution from a peaked central mass is remarkable, exceeding the density contribution from the disk and bulge.

For many spirals, including the Milky Way, the innermost region exhibits rapid rotation velocities, offering evidence for massive nuclear black holes. Consequently, orbital velocities in the center decrease rapidly from a velocity close to the speed of light down to 100–200 km s $^{-1}$ of the bulge and disk rotation velocities (🔍 Fig. 19-7). It has been established that the spheroidal component, e.g., bulge, is deeply coupled with the central black hole, as inferred from the tight correlation between the black hole mass and bulge luminosity (Magorrian et al. 1998; Kormendy 2011).

Infrared observations of proper motions of circumnuclear stars revealed that the Galactic Center nests a black hole with mass of $3.7 \times 10^6 M_\odot$ coinciding with Sgr A* (Genzel et al. 1997, 2000; Ghez et al. 1998, 2005). In the nearby galaxy, NGC 4258, water maser lines at 22 GHz are observed from a disk of radius 0.1 pc in Keplerian rotation, which is attributed to a massive black hole of a mass of $3.9 \times 10^7 M_\odot$ (Herrnstein et al. 1999; Miyoshi et al. 1995; Nakai et al. 1993). VLBI observations of the water maser lines have revealed a rapidly rotating nuclear torus of sub parsec scales in several nearby active galactic nuclei. There have been an increasing number of evidences for massive black holes in galactic nuclei (Melia 2011).

The existence of massive objects in the nuclei is without doubt related to nuclear activity such as via rapid gas inflow in the steep gravitational potential. However, the activity is not always activated by the dynamical structure. It is well established that the rotation curves, and

therefore, the fundamental mass distributions in the central regions are very similar to each other. Showing the same dynamical structure, some galaxies exhibit nuclear activity, while the others do not.

In fact, high-accuracy central rotation curves for starburst galaxies, Seyferts, LINERs, and galaxies with nuclear jets do not show any particular peculiarity in their mass distributions. The typical starburst galaxy M82 shows normal central rotation, except that its outer rotation is declining due to the truncation of dark halo. Even such a very active galaxy like NGC 5128 (Cen A) shows a normal rotation curve (van Gorkom et al. 1990). The radio lobe galaxy NGC 3079 has both strong nuclear activity and usual rotation properties.

It is likely that nuclear activity is triggered by local and temporary causes around central massive cores and black holes, while the underlying dynamical structure is stable and universal in any galaxies. Possible mechanism would be intermittent inflow of gas from a circumnuclear torus, to which the disk gas is continuously accumulating by secular angular momentum loss by a bar or spiral arms. Because the accreting gas would be clumpy, the torus will be distorted from axisymmetric potential, triggering noncircular streaming flow to cause shocks and angular momentum exchange among the clumps. Some clumps will lose angular momentum, and are accreted to the nucleus.

6 Summary

The methods to derive rotation curves of the Galaxy and spiral galaxies were described in this chapter, and explanations were given of the methods to calculate the mass distribution using the rotation curve. Observed rotation curves for the Milky Way as well as for nearby spiral galaxies were presented, and were used to derive the mass distributions. The rotation curve is the most fundamental tool to discuss the dynamics and mass distribution in the Galaxy and in disk galaxies, and there are various methods to derive rotation curves. A unified rotation curve of the Milky Way was obtained by integrating the current various observations as well those for nearby spiral galaxies (➤ [Figs. 19-5](#), ➤ [19-10](#) and ➤ [19-11](#)).

The mass distribution is determined by both dynamical and photometric methods. In this chapter, the dynamical method was described in detail. Compared to the statistical method, which assumes such parameters as the mass-to-luminosity ratio, the dynamical method was shown to be more direct to measure the mass including the dark matter and black holes. The dynamical method may be further categorized into two ways: One is the decomposition method, by which the rotation curve is fitted by a calculated one as summation of several mass components (➤ [Figs. 19-13](#) and ➤ [19-14](#)). The dynamical parameters of individual components are determined during the fitting process. This method is convenient to discuss separately the basic galactic structures, which are usually the central massive object, bulge, disk, and dark halo (➤ [Table 19-2](#)). Each of these components may have its own evolutionary and dynamical properties. Inconvenience of this method was that one has to assume a priori the functional forms for the mass components, which may not necessarily be selected uniquely by observations. Therefore, the results depend on the adopted models, functional forms, as well as on one's consideration on the galactic structure.

Another method is the *t* method, in which the mass distribution is directly calculated from the data of rotation velocity (➤ [Fig. 19-16](#) and ➤ [19-17](#)). This method does not employ any galactic models or functional forms, but straightly compute the mass distribution. The results can be compared with surface photometry to obtain the distribution of mass-to-luminosity

ratio by simply dividing the surface mass density by surface luminosity (● *Fig. 19-24*). The thus computed M/L ratio is found to generally increase monotonically toward the edges of spiral galaxies, indicating the direct evidence for dark matter halo. As discussed in the previous sections, these two dynamical methods, decomposition and direct methods, are complimentary, and the derived mass distributions are consistent with each other.

It was shown that the rotation characteristics of spiral galaxies are similar to each other, and so are the mass distributions. The dynamical back bones are, thus, universal from Sa to Sc galaxies, and the structures are almost identical among the galaxies in so far as the dynamical mass is concerned. The Milky Way exhibits equally the most typical universal characteristics.

The dark halo of the Galaxy was discussed in detail. It was shown that the NFW and Burkert profiles better represent the observations of the outermost rotation characteristics up to $R \sim 150$ kpc compared to the isothermal profile. It was also shown that the dark halo extends far out to the intracluster space almost a half way to M31, where the dark matter properly possessed by the Local Group dominates in order for the group to be gravitationally bound.

Cross-References

- [Bulges](#)
- [Dark Matter in the Galactic Dwarf Spheroidal Satellites](#)
- [Disks and Warps](#)
- [History of Dark Matter in Galaxies](#)
- [Kinematics and Dynamics in the Galaxy](#)
- [The Galactic Nucleus](#)
- [The HI Galaxy](#)
- [The Molecular Galaxy](#)

References

- | | |
|---|--|
| <p>Amram, P., Boulesteix, J., Marcelin, M., Balkowski, C., Cayatte, V., & Sullivan, W. T., III, 1995, <i>AAS</i>, 113:35</p> <p>Arimoto, N., Sofue, Y., & Tsujimoto, T. 1996, <i>PASJ</i>, 48, 275</p> <p>Ashman, K. M. 1992, <i>PASP</i>, ApJ, 104, 1109</p> <p>Athanassoula, E. 1992, <i>MNRAS</i>, 259, 345</p> <p>Bally, J., Stark, A. A., Wilson, R. W., & Henkel, C. 1987, <i>ApJS</i>, 65, 13</p> <p>Begeman, K. G. 1989, <i>AA</i>, 223, 47</p> <p>Begeman, K. G., Broeils, A. H., Sanders, R. H. 1991, <i>MNRAS</i>, 249, 523</p> <p>Bertola, F., Cappellari, M., Funes, J. G., Corsini, E. M., Pizzella, A., Vega Bertran, J. C. 1998, <i>ApJ Lett</i>, 509, 93</p> <p>Binney, J., Gerhard, O. E., Stark, A. A., Bally, J., & Uchida, K. I. 1991, <i>MNRAS</i>, 252, 210</p> <p>Binney, J., & Tremaine, S. 1987, in <i>Galactic Dynamics</i> (Princeton: Princeton Univ. Press)</p> | <p>Binney, J. 1982, <i>ARAA</i>, 20, 399</p> <p>Blais-Ouellette, S., Amram, P., & Carignan, C. 2001, <i>AJ</i>, 121, 1952</p> <p>Blitz, L., Fich, M., & Stark, A. A. 1982, <i>ApJs</i>, 49, 183</p> <p>Blitz, L., Binney, J., Lo, K. Y., Bally, J., & Ho, P. T. P. 1993, <i>Nature</i>, 361, 417</p> <p>Bosma, A. 1981a, <i>AJ</i>, 86, 1825</p> <p>Bosma, A. 1981b, <i>AJ</i>, 86, 1791</p> <p>Burbidge, E. M., Burbidge, G. R., Crampin, D. J., Rubin, V. C., Prendergast, K. H. 1964, <i>ApJ</i>, 139, 1058</p> <p>Burbidge, E. M., Burbidge, G. R. 1975, in <i>Stars and Stellar Systems IX: Galaxies and the Universe</i>, ed. A. Sandage, M. Sandage, & J. Kristian (Chicago, IL: Univ. Chicago Press), 81</p> <p>Bureau, M., & Athanassoula, E. 1999, <i>ApJ</i>, 522, 686</p> <p>Burkert, A. 1995, <i>ApJ</i>, 447, L25</p> <p>Burton, W. B., & Gordon, M. A. 1978, <i>AA</i>, 63, 7</p> <p>Carignan, C. 1985, <i>ApJ</i>, 299, 59</p> |
|---|--|

- Casertano, S., & van Gorkom, J. H. 1991, *AJ*, 101, 1231
- Ciotti, L. 1991, *AA*, 249, 99
- Clemens, D. P. 1985, *ApJ*, 295, 422
- Corbelli, E., & Salucci, P. *MNRAS*, 2000, 311, 441
- Courteau, S. 1997, *AJ*, 114, 2402
- Dame, T. M., Ungerechts, H., & Cohen, R. S., et al. 1987, *ApJ*, 322, 706
- de Blok, W. J. G. 2005, *ApJ*, 634, 227
- Deguchi, S., Fujii, T., Izumiura, H., Kameya, O., Nakada, Y., Nakashima, J., Ootsubo, T., & Ukita, N. 2000, *ApJ*, Suppl 128, 571
- Demers, S., & Battinelli, P. 2007, *AA*, 473, 143
- de Vaucouleurs, G. 1953, *MNRAS*, 113, 134
- de Vaucouleurs, G. 1958, *ApJ*, 128, 465
- de Zeeuw, T., & Franx, M. 1991, *ARAA*, 29, 239
- Einasto, J., Saar, E., Kaasik, A., & Chernin, A. D. 1974, *Nature*, 252, 111
- Faber, S. M., & Gallagher, J. S. 1979, *ARAA*, 17, 135
- Fich, M., Blitz, L., & Stark, A. 1989, *ApJ*, 342, 272
- Fich, M., & Tremaine, S. 1991, *ARAA*, 29, 409
- Forbes, D. A. 1992, *AAS*, 92, 583
- Freeman, K. C. 1970, *ApJ*, 160, 811
- Freudenreich, H. T. 1998, *ApJ*, 492, 495
- Genzel, R., Thatte, N., Krabbe, A., Kroker, H., & Tacconi-Garman, L. 1996, *ApJ*, 472, 153
- Genzel, R., Eckart, A., Ott, T., & Eisenhauer, F. 1997, *MNRAS*, 291, 219
- Genzel, R., Pichon, C., Eckart, A., Gerhard, O. E., & Ott, T. 2000, *MNRAS*, 317, 348
- Ghez, A., Morris, M., Klein, B. L., & Becklin, E. E. 1998, *ApJ*, 509, 678
- Ghez, A. M., Salim, S., Hornstein, S. D., Tanner, A., Lu, J. R., Morris, M., Becklin, E. E., & Duchene, G. 2005, *ApJ*, 620, 744
- Gilmore, G., King, I. R., & van der Kruit, P. C. 1990, in *The Milky Way as a Galaxy* (Mill Valley, CA: Univ. Science Books)
- Herrnstein, J. R., Moran, J. M., Greenhill, L. J., Diamond, P. J., Inoue, M., Nakai, N., Miyoshi, M., Henkel, C., & Riess, A. 1999, *Nature*, 400, 539
- Honma, M., Bushimata, T., Choi, Y. K., Hirota, T., & Imai, H. et al. 2007, *PASJ*, 59, 839
- Honma, M., & Sofue, Y. 1997, *PASJ*, 49, 453
- Irwin, J., Judith, A., & Seaquist, E. R. 1991, *ApJ*, 371, 111
- Izumiura, H., Deguchi, S., Fujii, T., Kameya, O., Matsumoto, S., Nakada, Y., Ootsubo, T., & Ukita, N. 1999, *Ap J Suppl*, 125, 257
- Jobin, M., & Carignan, C. 1990, *AJ*, 100, 648
- Kahn, F. D., & Woltjer, L. 1959, *ApJ*, 130, 705
- Kenney, J. D. P., Wilson, C. D., Scoville, N. Z., Devereux, N. A., & Young, J. S. 1992, *ApJL*, 395, L79
- Kent, S. M. 1986, *AJ*, 91, 1301
- Kent, S. M. 1992, *ApJ*, 387, 181
- Kent, S. M., Dame, T. M., & Fazio, G. 1991, *ApJ*, 378, 131
- Kong, D. L., & Zhu, Z. 2008, *Acta Astronomica Sinica*, 49, 224
- Kormendy, J. 2011, in this volume
- Kuno, N., Nishiyama, K., Nakai, N., Sorai, K., & Vila-Vilaro, B. 2000, *PASJ*, 52, 775
- Li, Y.-S., & White, S. D. M. 2008, *MNRAS*, 384, 1459
- Lindqvist, M., Habing, H. J., & Winnberg, A. 1992, *AA*, 259, 118
- Mathewson, D. S., Ford, V. L., & Buchhorn, M. 1992, *ApJ*, Suppl 81, 413
- Mathewson, D. S., & Ford, V. L. 1996, *ApJ*, Supp 107, 97
- Magorrian, J., et al. 1998, *AJ*, 115, 2285
- Melia, F. 2011, in this volume
- Merrifield, M. R. 1992, *AJ*, 103, 1552
- Miyamoto, M., & Nagai, R. 1975, *PASJ*, 27, 35
- Miyoshi, M., Moran, J., Herrnstein, J., Greenhill, L., Nakai, N., Diamond, P., & Inoue, M. 1995, *Nature*, 373, 127
- Nakai, N., Inoue, M., & Miyoshi, M. 1993, *Nature*, 361, 45
- Nakai, N. 1992, *PASJ*, 44, L27
- Nakanishi, H. 2007, in *Modern Astronomy Series* Vol. 5, "The Galaxy", Chap. 2, eds. Y. Sofue, N. Arimoto, & M. Iye (Tokyo: Nihon Hyoronsha. Co.)
- Nakanishi, H., & Sofue, Y. 2003, *PASJ*, 55, 191
- Nakanishi, H., & Sofue, Y. 2006, *PASJ*, 58, 847
- Navarro, J. F., Frenk, C. S., White, S. D. M., 1996, *ApJ*, 462, 563
- Noordermeer, E. 2008, *MNRAS*, 385, 1359
- Noordermeer, E., van der Hulst, J. M., Sancisi, R., Swaters, R. S., & van Albada, T. S. 2007, *MNRAS*, 376, 1513
- Oka, T., Hasegawa, T., Sato, F., Tsuboi, M., & Miyazaki, A. 1998, *ApJ*, Suppl 18, 455
- Oort, J. H. 1940, *ApJ*, 91, 273
- Oort, J. H. 1965, in *Stars and Stellar Systems*, Vol. 5, Galactic Structure, eds. A. Blaauw, & M. Schmidt (Chicago, IL: Univ. Chicago Press), 455
- Ostriker, J. P., & Peebles, P. J. E. 1973, *ApJ*, 186, 467
- Peebles, P. J. E. 1995, *ApJ*, 449, 52
- Persic, M., & Salucci, P. eds. 1997, *Dark Matter in the Universe* (San Francisco, CA: ASP)
- Persic, M., Salucci, P., & Stel, F. 1996, *MNRAS*, 281, 27
- Reid, M. J. 1993, *ARAA*, 31, 345
- Roberts, M. A., & Rots, A. H. 1973, *AA*, 26, 483
- Roelfsema, P. R., & Allen, R. J. 1985, *AA*, 146, 213
- Roscoe, D. F. 1999, *AA*, 343, 788
- Rubin, V. C., Burstein, D., Ford Jr W. K., & Thonnard, N. 1985, *ApJ*, 289, 81
- Rubin, V. C., & Ford Jr W. K. 1970, *ApJ*, 159, 379
- Rubin, V. C., & Ford Jr W. K. 1983, *ApJ*, 271, 556R

- Rubin, V. C., & Ford Jr W. K., & Thonnard, N. 1982, *ApJ*, 261, 439
- Rubin, V. C., & Graham, J. A. 1987, *ApJL*, 316, L67
- Rubin, V. C., Kenny, J. D. P., & Young, J. S. 1997, *AJ*, 113, 1250
- Rubin, V. C., Waterman, A. H., & Kenney, J. D. 1999, *AJ*, 118, 236
- Sakamoto, K., Okumura, S. K., Ishizuki, S., & Scoville, N. Z. 1999, *ApJS*, 124, 403
- Sargent, A. I., & Welch, W. J. 1993, *ARAA*, 31, 297
- Sawa, T., & Fujimoto, M. 2005, *PASJ*, 57, 429
- Sérsic J. L., 1968, *Atlas de Galaxias Australes* (Cordoba, Argentina: Observatorio Astronomico)
- Sofue, Y. 1995, *PASJ*, 47, 527
- Sofue, Y. 1996, *ApJ*, 458, 120
- Sofue, Y. 1997, *PASJ*, 49, 17
- Sofue, Y. 1998, *PASJ*, 50, 227
- Sofue, Y. 2000, *PASJ*, 51, 445
- Sofue, Y. 2009, *PASJ*, 61, 153
- Sofue, Y., Honma, M., Omodaka, T. 2009, *PASJ*, 61, 229
- Sofue, Y., Koda, J., Kohno, K., Okumura, S. K., Honma, M., Kawamura, A., & Irwin, J. A. 2001, *ApJL*, 547, L115
- Sofue, Y., & Rubin, V. C. 2001, *ARAA*, 39, 137
- Sofue, Y., Tutui, Y., Honma, M., Tomita, A., Takamiya, T., Koda, J., & Takeda, Y. 1999, *ApJ*, 523, 136
- Spergel, D. N., et al. 2003, *ApJs*, 148, 175
- Spitzer, L. 1942, *ApJ*, 95, 329
- Swaters, R. A., Sancisi, R., van Albada, T. S., & van der Hulst, J. M. 2009, *AA*, 493, 871
- Takamiya, T., & Sofue, Y. 2000, *ApJ*, 534, 670
- Takamiya, T., & Sofue, Y. 2002, *ApJ*, 576, L15
- Toomre, A., & Toomre, J. 1972, *ApJ*, 178, 623
- Trimble, V. 1987, *ARAA*, 25, 425
- Trujillo, I., et al. 2002, *MNRAS*, 333, 510
- Tully, R. B., & Fisher, J. R. 1977, *AA*, 54, 661.
- van Albada, T. S., Bahcall, J. N., Begeman, K., & Sancisi, R. 1985, *ApJ*, 295, 305
- van Gorkom, J. H., van der Hulst, J. M., Haschick, A. D., & Tubbs, A. D. 1990, *AJ*, 99, 1781
- van der Kruit, P. C., & Allen, R. J. 1978, *ARAA*, 16, 103
- van der Marel, R. P., & Guhathakurta, P. 2008, *ApJ*, 678, 187
- Wada, K., & Habe, A. 1995, *MNRAS*, 277, 433
- Weiland, J. L., Arendt, R. G., Berriman, G. B., Dwek, E., Freudenreich, H. T., et al. 1994, *ApJ*, 425, L81
- Weiner, B., & Sellwood, J. A. 1999, *ApJ*, 524, 112
- Wilkinson, M. I., & Evans, N. W. 1999, *MNRAS*, 310, 645
- Wozniak, H., & Pfenniger, D. 1997, *AA*, 317, 14
- Zaritsky, D., Smith, R., Frenk, C., & White, S. D. M. 1997, *ApJ*, 478, 3

20 Dark Matter in the Galactic Dwarf Spheroidal Satellites

Matthew Walker

Harvard-Smithsonian Center for Astrophysics, Cambridge,
MA, USA
Hubble Fellow

1	<i>Introduction</i>	1041
2	<i>Observations</i>	1041
2.1	Census	1042
2.2	Stellar Structure	1044
2.2.1	“Classical” dSphs	1044
2.2.2	Ultrafaint dSphs	1047
2.2.3	Extended Structure	1049
2.2.4	Structural Peculiarities of Individual dSphs	1049
2.3	Stellar Velocities	1052
2.3.1	Small Number Statistics	1052
2.3.2	Confirmation	1053
2.3.3	Large Samples	1054
2.3.4	(Necessarily) Small Samples	1056
2.4	The Smallest Galaxies	1058
3	<i>Stellar Velocity Dispersion as a Proxy for Mass</i>	1060
3.1	Rotation	1060
3.2	External Tides	1060
3.3	Binary Stars	1062
4	<i>dSph Masses</i>	1064
4.1	Amount	1064
4.1.1	“Mass-Follows-Light” Models	1064
4.1.2	Does Mass Follow Light?	1065
4.1.3	Jeans Analysis	1067
4.2	Distribution	1070
4.2.1	Indirect Constraints	1071
4.2.2	Constraints from Models	1071
4.2.3	Direct Measurement	1072
5	<i>A Common dSph Mass?</i>	1072
6	<i>Implications for Cosmology and Particle Physics</i>	1075

6.1	Cosmology	1076
6.2	Particle Physics	1078
7	<i>Some Considerations for Future Work</i>	1081
	<i>References</i>	1082

Abstract: The Milky Way’s dwarf spheroidal satellites include the nearest, smallest, and least luminous galaxies known. They also exhibit the largest discrepancies between dynamical and luminous masses. This article reviews the development of empirical constraints on the structure and kinematics of dSph stellar populations and discusses how this phenomenology translates into constraints on the amount and distribution of dark matter within dSphs. Some implications for cosmology and the particle nature of dark matter are discussed and some avenues for future study are identified.

1 Introduction

Physics assigns to gravity the responsibility of forming structure on scales ranging from terrestrial to cosmological. An apparent threshold arises somewhere between scales characteristic of star clusters and those characteristic of galaxies. The internal dynamics of gravitationally bound structures smaller than a few parsecs, of which globular clusters are the largest examples, are reasonably well described in terms of standard gravity (i.e., Einstein’s general theory and/or its Newtonian approximation) sourced by known substances. The internal dynamics of gravitationally bound structures larger than a few tens of parsecs, of which dwarf spheroidal (dSph) galaxies are the smallest examples, are not.

The ubiquity of dark matter on galactic and larger scales signifies new physics. Either there exists an otherwise unknown substance that contributes to the dynamical mass but not to the baryonic mass, or the standard dynamical framework requires modification (or both). The “substance” hypothesis is not falsifiable, but in principle it can be confirmed with the detection of nongravitational interactions involving dark matter particles. In any case, dSphs provide the most extreme examples of dark matter phenomenology, with dynamical mass-to-light ratios $M/L_V \gtrsim 10 [M/L_V]_{\odot}$ even at their centers. This fact has made dSphs the focus of intense scrutiny in the effort to understand the nature of dark matter.

This article reviews the development of empirical constraints on the amount and distribution of dark matter within the Milky Way’s dSph satellites. These results follow from the application of a rich variety of analyses applied to observations conducted by many individuals and groups. All analyses described here are formulated within the Newtonian dynamical framework. The reader is welcome to interpret “dark matter” in terms of the substance hypothesis or more generally as a quantification of the discrepancy between dynamical and baryonic mass. The focus here is on the relationships between data and constraint, and one expects that this information can be translated meaningfully into alternative dynamical frameworks (as formulated, e.g., by Bekenstein 2004; Bergmann 1968; Milgrom 1983; Moffat 2006).

2 Observations

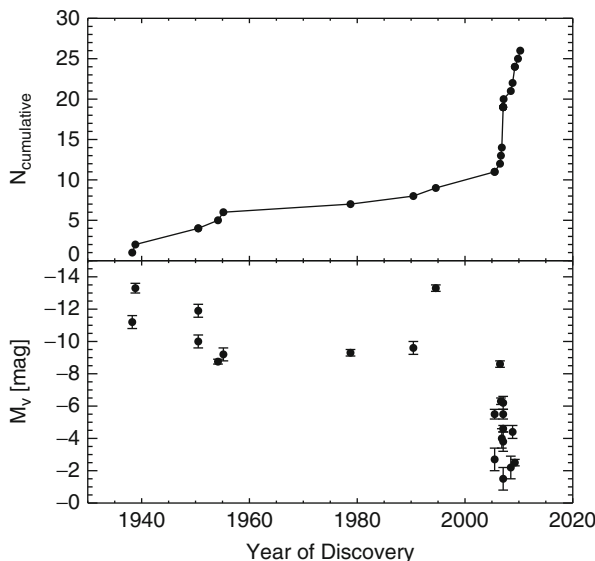
Bright red giant stars are detectable as point sources with $m_V \lesssim 21$ mag out to distances of ~ 0.5 Mpc. Within this range, the Milky Way’s dSph galactic satellites appear as localized overdensities of individually resolved stars. Empirical information about the number, stellar structure and internal dynamics of these systems has steadily accumulated for the past eight decades.

2.1 Census

Shapley (1938a) discovered the Sculptor dSph upon visual examination of a photographic plate exposed for 3 h with the 24-in. Bruce telescope at (what was then) Harvard's Boyden observatory in South Africa. In this original dSph discovery paper, Shapley notes that Sculptor is visible – in hindsight – as a faint patch of light on a plate taken for the purpose of site testing, in 1908, during a series of five exposures totalling nearly 24 h with a 1-in. refracting telescope.

Reporting shortly thereafter the similar discovery of Fornax, Shapley (1938b) notes that the new type of “Sculptor-Fornax” cluster shares properties with both globular clusters and elliptical galaxies, then speculates that “At the distance of the Andromeda system these objects would, in fact, have long escaped discovery. There may be several others in the Local Group of galaxies; such objects may be of frequent occurrence in intergalactic space and of much significance both in the census and the genealogy of sidereal systems.”

► *Figure 20-1* demonstrates Shapley's prescience, plotting the cumulative number and luminosities of the Milky Way's known dSph satellites against date of discovery publication. Harrington and Wilson (1950) and Wilson (1955) found the next four “Sculptor-type” (as they were then called) systems – Leo I, Leo II, Draco, and Ursa Minor – on photographic plates taken for the Palomar Observatory Sky Survey with the 48-in. Schmidt telescope. Cannon et al. (1977) spotted Carina by eye on a plate taken with the 1.2-m UK-Schmidt telescope. Irwin et al. (1990) used the automated photographic measuring (APM) facility at the University of Cambridge to




■ Fig. 20-1


Cumulative number (*top*) and luminosity (*bottom*) of known dwarf spheroidal satellites of the Milky Way versus year of publication of discovery paper (Belokurov et al. 2006a, 2007, 2008, 2009, 2010; Cannon et al. 1977; Grillmair 2009; Harrington and Wilson 1950; Ibata et al. 1994; Irwin et al. 1990, 2007; Shapley 1938a, b; Walsh et al. 2007; Watkins et al. 2009; Willman et al. 2005a, b; Wilson 1955; Zucker et al. 2006a, b)

detect Sextans, again on a UK-Schmidt plate, thereby completing the census of the Milky Way's eight so-called "classical" dSphs.

The next discovery was unique in that it came via spectroscopy rather than imaging. During a stellar kinematic survey of the Galactic bulge, Ibata et al. (1994) noticed Sagittarius (Sgr) in color-velocity space as an excess of stars following a narrow velocity distribution offset from that of bulge members. Subsequent observations showed that Sgr, located just ~ 15 kpc behind the Galactic center, spews stars along tidal tails that wrap spectacularly around the entire sky (e.g., Belokurov et al. 2006b; Ibata et al. 1997; Koposov et al. 2011a; Majewski et al. 2003; Mateo et al. 1996). The tidal disruption of Sgr provides a useful tracer of the Galactic potential (e.g., Johnston et al. 2005; Law and Majewski 2010; Peñarrubia et al. 2010), but disqualifies Sgr from a simple equilibrium-dynamical analysis. For this reason, Sgr shall not be considered further here.

In a span of 7 years, deep multicolor photometry from the Sloan Digital Sky Survey (SDSS) has now tripled the number of known Milky Way satellites (Belokurov et al. 2006a, 2007, 2008, 2009, 2010; Grillmair 2009; Irwin et al. 2007; Walsh et al. 2007; Watkins et al. 2009; Willman et al. 2005a, b; Zucker et al. 2006a, b). The 17 satellites discovered with SDSS data have extended the floor of the observed galaxy luminosity function from $M_V \gtrsim -8$ to $M_V \gtrsim -2$ (bottom panel of  Fig. 20-1), such that some galaxies are less luminous than some stars! Unlike their brighter "classical" siblings, the "ultrafaint" satellites¹ discovered with SDSS data are not apparent to the eye, even in deep images. Rather, they are detected only by correlating spatial overdensities with overdensities in color-magnitude space (e.g., Belokurov et al. 2006a; Walsh et al. 2008). In order to confirm the faintest satellites, the SDSS catalog must be supplemented by deeper, follow-up photometry as well as spectroscopy (e.g., Belokurov et al. 2009, 2010). Given the rate of false positives expected for candidates remaining in SDSS data ($\gtrsim 4/5$ in the author's experience), the expense of follow-up observations can be prohibitive. However, the next generation of sky surveys (SkyMapper, Pan-STARRS, DES, Gaia, LSST, etc.) will almost certainly bring a new flurry of discoveries, particularly in the relatively unexplored southern sky.

Although they are not considered further below, it is worth noting that Shapley was correct regarding the dSph satellites of M31. van den Bergh (1972) discovered the first example with the Palomar Schmidt telescope using plates more sensitive than those used in the original Palomar Survey. The current census includes 27 known dSph satellites of M31. Two-thirds of this number were discovered in the past 7 years (Bell et al. 2011; Ibata et al. 2007; Irwin et al. 2008; Majewski et al. 2007; Martin et al. 2006, 2009; McConnachie et al. 2008; Richardson et al. 2011; Slater et al. 2011; Zucker et al. 2004, 2007), with SDSS data as well as photometry from the PAndAS survey conducted with the Canada-France-Hawaii Telescope (McConnachie et al. 2009).

¹As  Fig. 20-1 and the terms themselves suggest, the distinction between "classical" and "ultrafaint" dSphs involves a mixture of intrinsic luminosity with sequence of discovery. Here, this distinction (which is meaningless in the sense that members of both classes trace smooth scaling relationships involving luminosity, size, metallicity, and stellar kinematics) is preserved only because observational studies of these objects – for both practical and accidental reasons – tend to be separable along the same lines. Following common practice, dSphs known before SDSS (Carina, Draco, Fornax, Leo I, Leo II, Sculptor, Sextans, Ursa Minor) are referred to as "classical," and the rest as "ultrafaint."

2.2 Stellar Structure

Galactic dynamics is concerned with the relationship between the gravitational potential and the distribution of stars in phase space. Observers who study dark matter in dSph galaxies must gather information about the positions and velocities of dSph stars. The largest dSphs subtend solid angles of several square degrees, making it difficult to study their stellar structure at large radius. Complete homogenous surveys are rare and valuable.

2.2.1 “Classical” dSphs

Hodge (1961a, b, 1962, 1963, 1964a, b) used photographic plates obtained at the Palomar (48-in. Schmidt, 100- and 200-in.), Lick (120-in.), and Boyden Station (24-in. Schmidt) in order to study luminous structure of the six Milky Way dSphs known at the time (Sculptor, Fornax, Leo I, Leo II, Draco, and Ursa Minor). Hodge counted stars within squares of regular grids overlaid on each plate and provided this description of analogue data reduction: “Each plate was counted at one sitting so that uniformity would be maintained. The plates were counted to the limiting magnitude and each was counted once. From experience ... it was decided that the reproducibility of counts on a particular plate is greater than from plate to plate, so that it is better to count many plates each once than one plate many times” (Hodge 1961b).

► *Figure 20-2* (top two rows of panels) displays isopleth maps that Hodge drew by connecting squares containing equal numbers of stars. These maps reveal that the internal structure of dSphs is smooth. Hodge reasoned that a well-mixed dSph requires dynamical relaxation by a process other than stellar encounters as the low surface densities of dSphs imply internal relaxation timescales of $\gtrsim 10^3$ Hubble times. Hodge (1966) and Hodge and Michie (1969) would later suggest that encounters between stellar groups might enable exchanges of orbital energy over shorter timescales during dSph formation. Eventually, Lynden-Bell (1967) would show that the time-varying gravitational potential of a young galaxy effectively shuffles the orbital energies of its stars, generating “violent” relaxation without stellar encounters. More recently, Mayer et al. (2001b) have proposed a mechanism specific to dSphs, demonstrating with N-body+hydrodynamical simulations that repeated tidal encounters with the Milky Way can effectively transform a rotating dSph progenitor into a pressure-supported spheroid in less than a Hubble time.

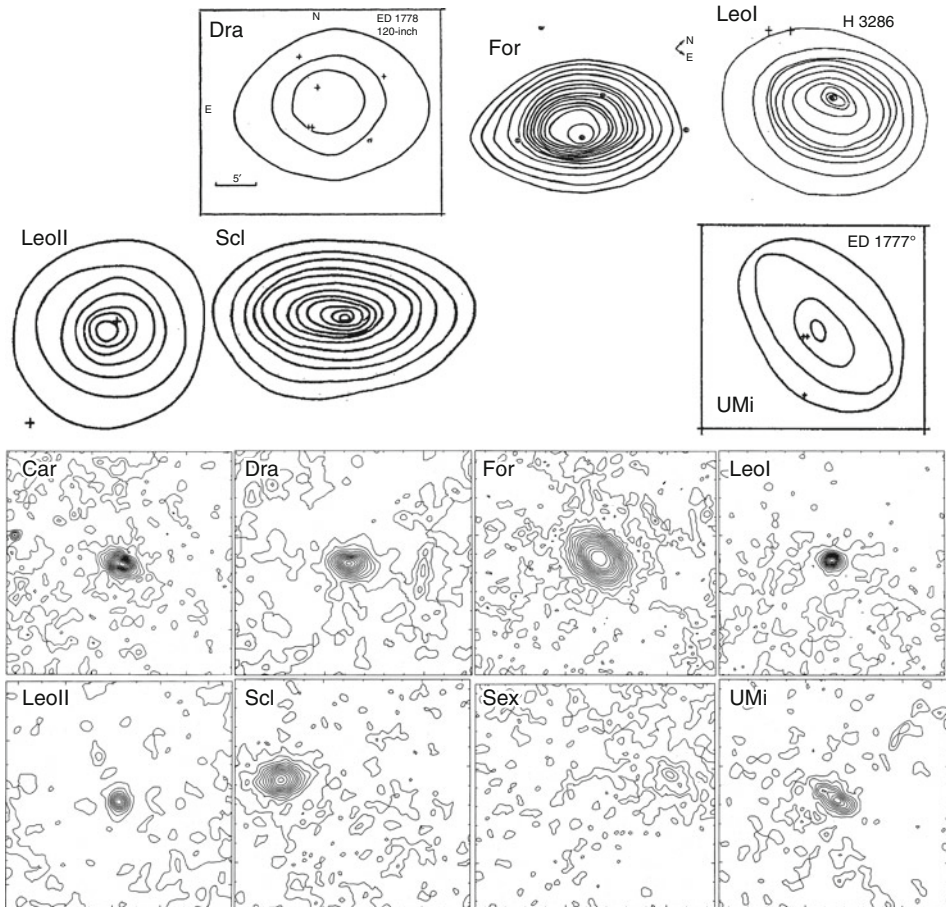
Hodge discovered two other structural features common to the classical dSphs. First, most exhibit flattened morphology, with typical ellipticities of $\epsilon \equiv 1 - b/a \sim 0.3$, where a and b are semimajor and semiminor axes, respectively. Second, dSph stellar density profiles decline more steeply at large radius than do the profiles of giant elliptical galaxies. Whereas the latter are commonly fit by formulae with relatively shallow outer profiles, e.g., $\Sigma(R) = \Sigma(0)/(1 + R/a)^2$ (Hubble 1930) or $\Sigma(R) = \Sigma(0) \exp[-kR^{1/4}]$ (de Vaucouleurs 1948), Hodge found that classical dSphs all have steeper outer profiles that are better fit with the formula of King (1962):

$$\Sigma(R) = k \left[\frac{1}{\sqrt{1 + (R/R_c)^2}} - \frac{1}{\sqrt{1 + (R_K/R_c)^2}} \right]^2, \quad (20.1)$$

where R_c is a “core” radius and R_K is a maximum, or “limiting” radius that one might expect to result from tidal truncation (► Sect. 2.2.3).²

²In fact R_K is usually referred to as a “tidal” radius and denoted r_t . The adopted nomenclature and notation avoid confusion with the tidal radius defined in (► 20.4).

Three decades later, Irwin and Hatzidimitriou (1995, “IH95” hereafter) used the APM facility at Cambridge to count stars automatically on photographic plates from the Palomar and UK-Schmidt telescopes. While confirming Hodge’s findings, IH95 produced significantly deeper maps and were able to include Carina and Sextans, the two Milky Way dSphs discovered in the interim (► Fig. 20-2). IH95 used these maps to measure the centroid, ellipticity, and orientation of each dSph, and then to tabulate stellar density as a function of distance along the semimajor axis. While the homogenous analysis of IH95 continues to provide a valuable resource particularly for comparing dSphs in the context of scaling relations (► Sect. 5), deeper photometric data sets now exist for most of the classical dSphs (e.g., Battaglia et al. 2006; Coleman et al. 2005a, a; Coleman et al. 2005b, b; Lee et al. 2003; Majewski et al. 2000;



■ Fig. 20-2

Stellar isodensity maps for the Milky Way’s eight “classical” dSphs. *Top two rows*: from Hodge’s star count studies (Hodge 1961a, b, 1962, 1963, 1964a, b, reproduced by permission of the American Astronomical Society). *Bottom two rows* (Reproduced from *Structural Parameters for the Galactic Dwarf Spheroidals*, by Irwin and Hatzidimitriou (1995), by permission of Wiley)


Odenkirchen et al. 2001; Palma et al. 2003; Saviane et al. 2000; Stetson et al. 1998; Tolstoy et al. 2004; Walcher et al. 2003; Westfall et al. 2006).

In principle the structure of dSph stellar components carries information about the mechanisms that drive dSph formation and evolution. While incompatible with shallow Hubble and de Vaucouleurs profiles, the available data often do not distinguish the King profile (Equation 1) from other commonly adopted fitting formulae—e.g., exponential and Plummer (1911) profiles,

$$\Sigma(R) = \Sigma(0) \exp[-R/R_e] \quad (20.2)$$

and

$$\Sigma(R) = \frac{\Sigma(0)}{[1 + (R/R_h)^2]^2}, \quad (20.3)$$

respectively, where $R_h \approx 1.68R_e$ is the projected half-light radius (i.e., the radius of the circle enclosing half the stars as viewed in projection).  Figure 20-3 displays IH95's fits of King (1966) and exponential surface brightness profiles.

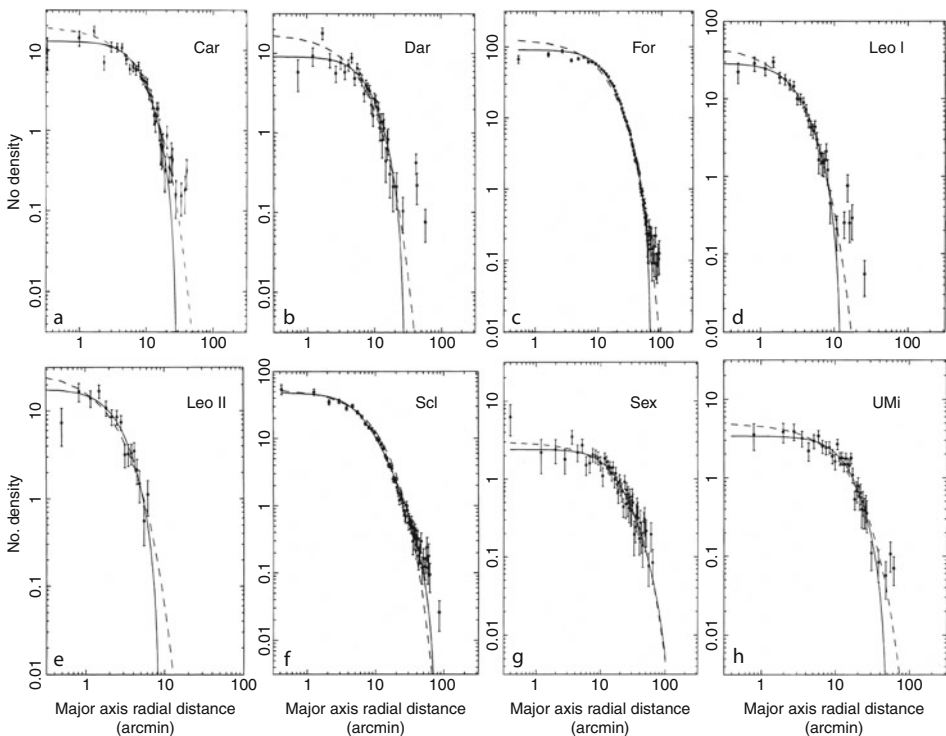
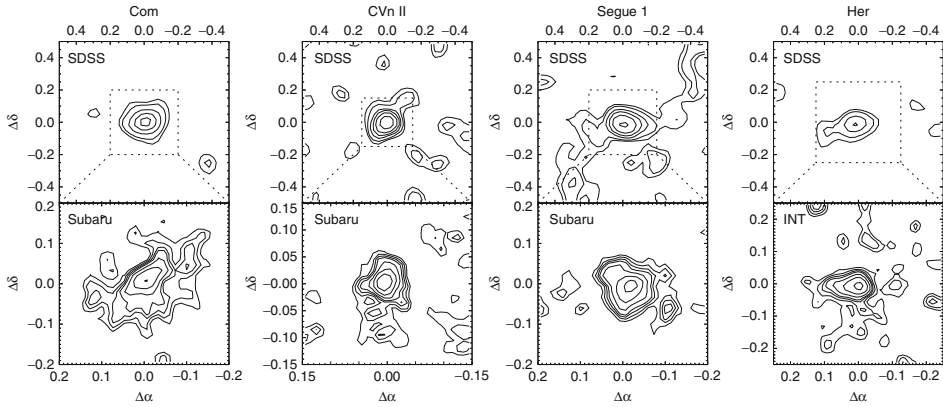


 Fig. 20-3

Stellar density profiles for the Milky Way's "classical" dSphs. Overlaid are best-fitting King (1966, solid) and exponential (dashed) surface brightness profiles (Reproduced from *Structural Parameters for the Galactic Dwarf Spheroidals* by Irwin and Hatzidimitriou (1995), by permission of Wiley)



■ Fig. 20-4

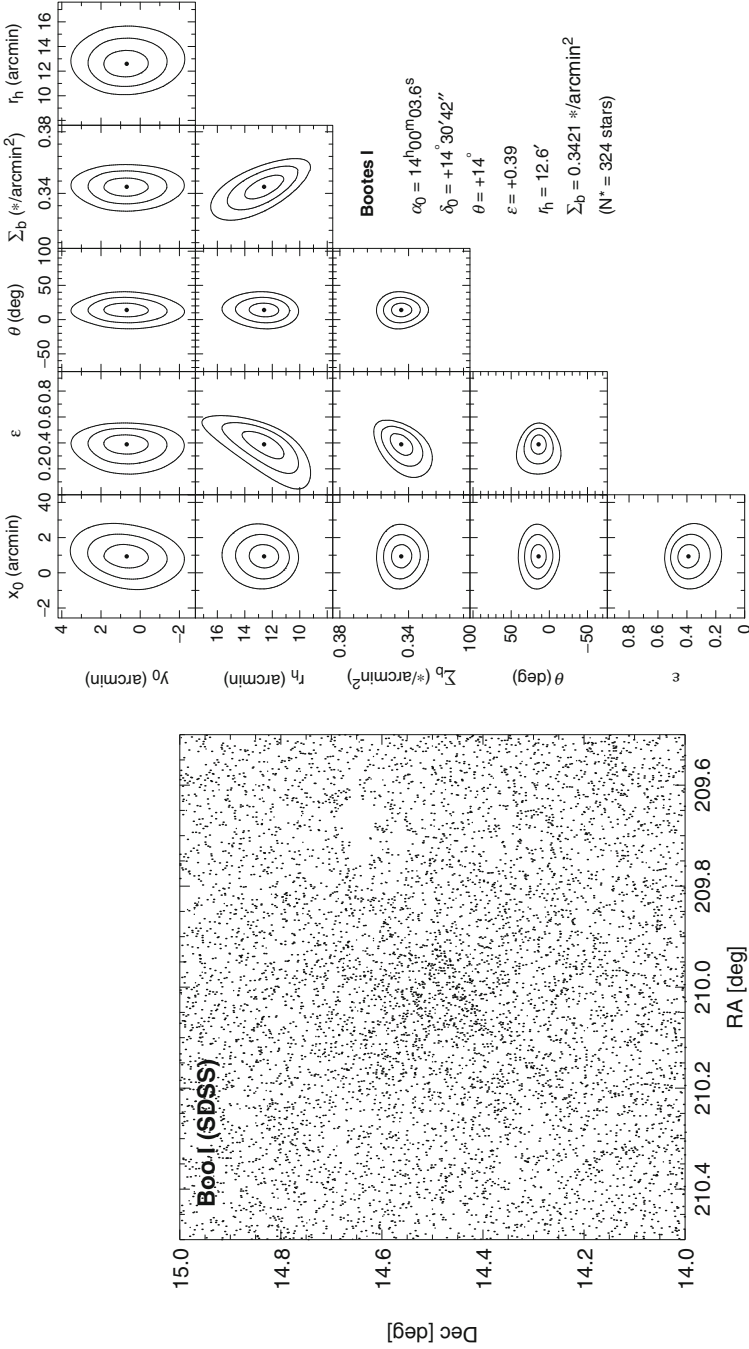
Stellar isodensity maps for four of the Milky Way’s “ultrafaint” dSphs from the discovery paper of Belokurov et al. (2007, reproduced by permission of the American Astronomical Society)

2.2.2 Ultrafaint dSphs

The SDSS catalog enables homogeneous studies of the structural properties of ultrafaint dSphs. For example, ● Fig. 20-4 displays isopleth maps constructed by Belokurov et al. (2007) for Coma Berenices, Canes Venatici II, Segue 1, and Hercules using both SDSS and deeper follow-up data. Whereas Hodge and IH95 estimated structural parameters for the classical dSphs after binning their star count data and subtracting estimates of foreground densities, the low surface brightnesses of “ultrafaint” dSphs are conducive neither to binning nor to foreground subtraction. Fortunately, neither procedure is necessary; indeed, Kleyna et al. (1998) estimate structural parameters for Ursa Minor using a likelihood function that specifies the probability associated with each individual stellar position in terms of a parametric surface brightness profile plus constant foreground.

Martin et al. (2008) design a similar maximum-likelihood analysis that operates directly on unbinned SDSS data in order to estimate the centroid, half-light radius, luminosity, ellipticity, and orientation of each ultrafaint satellite. ● Figure 20-5 demonstrates the efficacy of their method. The left panel maps individual stars from the SDSS catalog that are near the line of sight to Boötes I and satisfy color/magnitude criteria designed to select red giants at the distance of Boötes I. Panels on the right-hand side show maximum-likelihood estimates of each free parameter. Marginalised error distributions include the effects of sampling errors and parameter covariances and can be used directly in subsequent kinematic/dynamical analyses (● Sect. 4).

Martin et al. (2008) show that their method recovers robust estimates even when the SDSS sample includes as few as tens of satellite members. Subsequent tests with synthetic data by Muñoz et al. (2011) provide reason for caution, suggesting that for objects with low surface brightness, insufficient contrast between members and foreground can generate biased estimates of structural parameters. It is reassuring that deeper observations with CFHT



■ **Fig. 20-5** Measurement of structural parameters for the Boötes I dSph from SDSS data (Martin et al. 2008). *Left:* sky positions of red giant candidates selected from the SDSS catalog. *Right:* constraints on structural parameters from the maximum-likelihood analysis of Martin et al. (2008), reproduced by permission of the American Astronomical Society)

(Muñoz et al. 2010) and Subaru (Okamoto et al. 2012) yield structural parameters for subsets of ultrafaint satellites that agree well with the SDSS-derived estimates of Martin et al. (2008).

2.2.3 Extended Structure

The outer stellar structure of a given dSph is determined by some combination of formation processes and the subsequent evolution within the external potential of the Milky Way. In his structural analyses of the six dSphs known in the 1960s, Hodge compared the observed limiting radii, R_K (☛ 20.1), to simple estimates of “tidal” radii, r_t , beyond which stars escape into the external potential of the Milky Way (King 1966; von Hoerner 1957):

$$r_t = R_D \left[\frac{M_D}{(3 + e)M_{MW}} \right]^{1/3}. \quad (20.4)$$

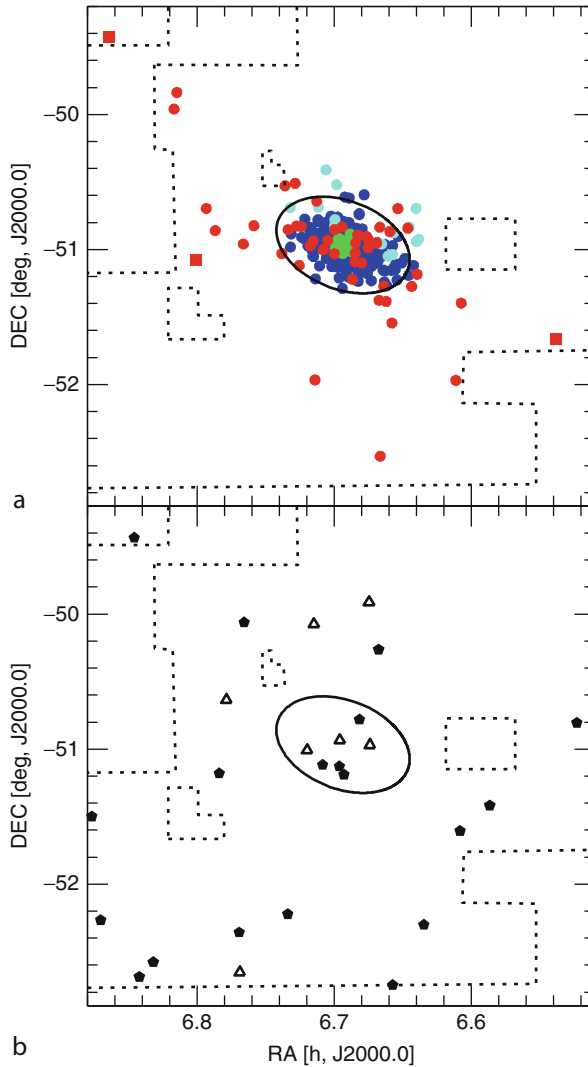
Here, M_{MW} and M_D are the Milky Way and dSph (point) masses, respectively, R_D is the pericentric distance of the dSph’s orbit and e is the orbital eccentricity. Assuming circular orbits and considering only luminous masses, Hodge (1966) noticed that while the two radii are similar for the nearest dSphs (Draco, Sculptor, Ursa Minor), for the three most distant dSphs (Leo I, Leo II, Fornax), $R_K \lesssim 0.5r_t$. Hodge (1966) concluded that tidal forces play a significant role in shaping the outer structures of the nearest dSphs.

Low surface brightness in the outer regions of dSphs makes extended structural studies difficult to conduct and interpret. For example, Coleman et al. (2005a) use deep wide-field photometry to conclude that Sculptor’s surface brightness profile is well described as the superposition of two equilibrium stellar components with different scale radii. Using an independent data set, Westfall et al. (2006) fit a single profile and identify a “break” in Sculptor’s surface brightness profile, which makes an apparent transition from a King profile in the inner parts to a shallower power law in the outer parts. Westfall et al. (2006) cite this transition not as the presence of a second component but rather as a signal that Sculptor is losing stars to tidal disruption.

In many cases, the use of narrow-band filters that are sensitive to stellar surface gravity (e.g., DDO51) can help to distinguish dSph red giants from foreground dwarf stars, providing a valuable boost in contrast (e.g., Majewski et al. 2000, 2005; Palma et al. 2003; Westfall et al. 2006). Having used narrow-band photometry to select spectroscopic targets at large dSph radii, Muñoz et al. (2005, 2006b) present velocities that confirm the membership of stars in Ursa Minor and Carina (☛ Fig. 20-6) out to radii ~ 3 and ~ 5 times larger, respectively, than estimates of R_K . These detections imply that tides can play a significant role in shaping the outer parts of dSphs (☛ Sect. 3.2).

2.2.4 Structural Peculiarities of Individual dSphs

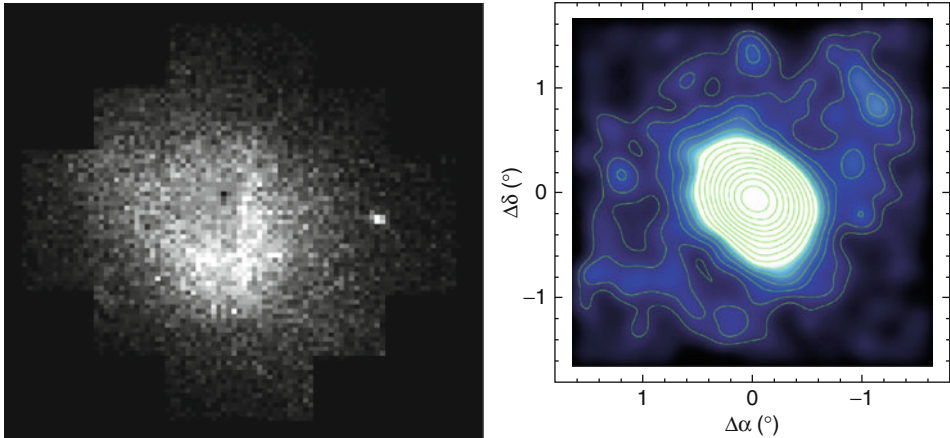
Most kinematic analyses of dSphs proceed from the assumption that dSphs host a single, spherically symmetric stellar component in dynamic equilibrium (☛ Sect. 4.1). However, real dSphs are all flattened (with $0.1 \lesssim \epsilon \equiv 1 - b/a \lesssim 0.7$, Irwin and Hatzidimitriou 1995; Martin et al. 2008; Sand et al. 2011), and it is not clear how severely this violation of spherical symmetry affects conclusions regarding dSph dynamics. In fact, most dSphs exhibit individual peculiarities that further violate the simplistic assumptions (☛ Sect. 4) employed in kinematic analyses.



■ Fig. 20-6

Extended stellar structure of the Carina dSph from narrow-band photometry and spectroscopy of Muñoz et al. (2006b, reproduced by permission of the American Astronomical Society). The *top/bottom panels* show sky positions of red giant candidates confirmed as Carina members/nonmembers. Ellipses mark Carina's limiting radius, R_K (● 20.1), determined from smooth fits to star count data. The extension of faint stellar structure at $R \gtrsim 3R_K$ suggests tidal interaction (● Sect. 3.2). Muñoz et al. (2005) report similar results for Ursa Minor

For example, Sculptor, Fornax, and Sextans all display evidence for chemodynamically independent stellar *sub*populations (Battaglia et al. 2006, 2011; Tolstoy et al. 2004, respectively). In all three cases, a relatively metal-rich, kinematically cold population has smaller scale radius than does a metal-poor, kinematically hot population (● Sect. 4.2.2, ● Fig. 20-10). Fornax also has irregular stellar structure in the form of a crescent-shaped feature near



■ Fig. 20-7

Structural irregularities in Fornax. *Left*: Gray scale map ($\sim 42'$ per side) of surface density of stars with $18.5 \leq V \leq 23$ from the photometry of Stetson et al. (1998, reproduced with permission from The University of Chicago Press). Notice the crescent-like feature near the center. *Right*: Smoothed stellar map from V, I photometry of Coleman et al. (2005b, reproduced by permission of the American Astronomical Society), who detect two shell-like features (one is visible ~ 1.3 deg northwest of the center) aligned with lobes along the morphological minor axis

its center (Stetson et al. 1998, and ▶ Fig. 20-7, left panel) and two shell-like features (Coleman et al. 2004) and lobes along its morphological minor axis (Coleman et al. 2005b and ◉ Fig. 20-7, right panel). These features suggest that Fornax may have undergone a recent merger, an event that might be related to the presence of a young (age ~ 100 Myr), centrally concentrated main sequence in Fornax (Battaglia et al. 2006). Ursa Minor exhibits clumpy stellar substructure, most dramatically in the form of a secondary peak in its luminosity distribution, offset by $\sim 20'$ from the central peak (Olszewski and Aaronson 1985). The region near the secondary peak is kinematically colder than the rest of Ursa Minor (Kleyna et al. 2003), and so may represent a bound star cluster (▶ Sect. 4.2.1).

Such peculiarities extend to the ultrafaint satellites as well. Spectroscopic surveys reveal that Segue 1 is superimposed on at least one stream of stellar debris (Geha et al. 2009; Niederste-Ostholt et al. 2009; Simon et al. 2011), offset by $\sim 100 \text{ km s}^{-1}$ from Segue 1 in velocity space. Segue 1 also shows hints of extended stellar structure superimposed on Sgr debris, although interpretation of these extremely low-surface-brightness features in terms of tidal disruption remains controversial (Belokurov et al. 2007; Geha et al. 2009; Niederste-Ostholt et al. 2009; Simon et al. 2011). Segue 2 appears to be embedded in – and comoving with – a stream of stellar debris, perhaps from a tidally disrupted parent system (Belokurov et al. 2009). Ursa Major II has flattened morphology and distorted outer morphology that suggest ongoing tidal disruption (Muñoz et al. 2010; Zucker et al. 2006a). Another candidate for disruption is Boötes III, which has irregular, clumpy morphology and a large measured velocity dispersion of $\sigma \sim 14 \text{ km s}^{-1}$ (Carlin et al. 2009). The alignment of stellar distortions in both Leo IV and Leo V hint at a low-surface-brightness “bridge” spanning the $\sim 20 \text{ kpc}$ between these systems (Belokurov et al. 2008; de Jong et al. 2010; Walker et al. 2009a). Stars nearest the center of the Willman 1 satellite

exhibit near-zero velocity dispersion and a mean velocity that is offset from the rest of Willman 1 members by $\sim 8 \text{ km s}^{-1}$ (Willman et al. 2011). Each newly discovered galaxy seems to present a new quirk of its own.

2.3 Stellar Velocities

Assuming that dSphs are purely stellar systems truncated by tidal interaction with the Milky Way, such that $r_t \sim R_K$, Ostriker et al. (1974) applied (20.4) to estimate the mass of the Milky Way. Inverting the calculation by assuming an isothermal Galactic halo with $v_{\text{circ}} = 225 \text{ km s}^{-1}$, Faber and Lin (1983) estimated masses of dSphs. For the nearest dSphs (Draco, Ursa Minor, Carina, Sculptor), Faber and Lin estimated mass-to-light ratios of $M/L_V \gtrsim 10[M/L_V]_{\odot}$, suggesting dSph dark matter. Faber and Lin further used their estimates to predict, via the virial theorem, values of $\gtrsim 10 \text{ km s}^{-1}$ for the internal stellar velocity dispersions of dSphs.

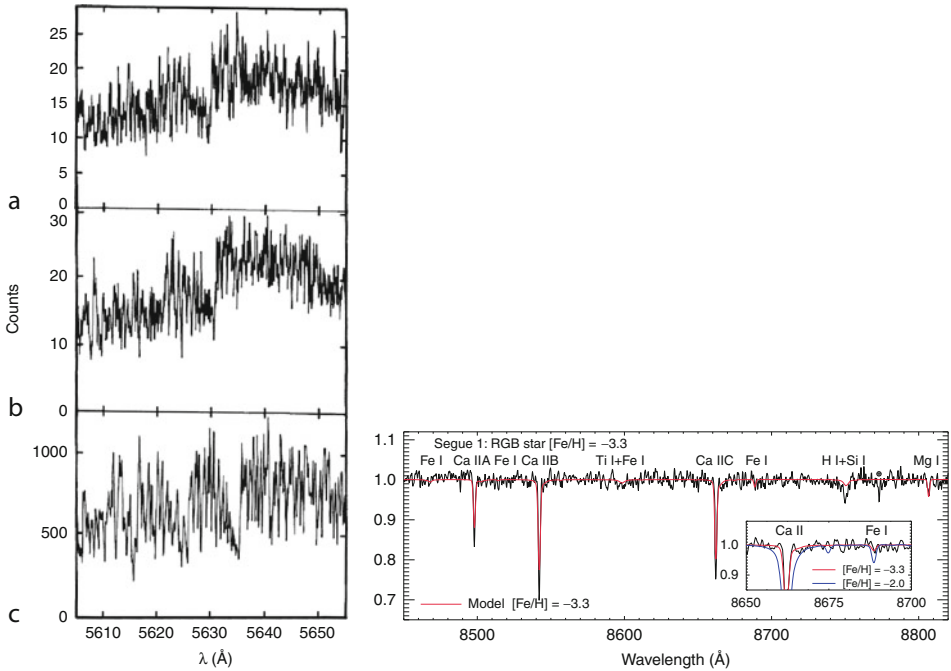
At the same time, Aaronson (1983) provided the first actual measurement of a dSph's internal velocity dispersion. Aaronson used the Multiple Mirror Telescope (MMT, which then consisted of six 1.8-m mirrors working in tandem) to acquire high-resolution ($R \sim 30,000$) spectra for three individual carbon stars in Draco. Figure 20-8 displays the spectra, which were sufficient for Aaronson to measure precise velocities of $-298.7 \pm 0.9 \text{ km s}^{-1}$ (with a follow-up measurement of $-297.6 \pm 0.6 \text{ km s}^{-1}$), $-300.2 \pm 0.6 \text{ km s}^{-1}$, and -279.7 km s^{-1} . Aaronson calculated that such measurements require, at the 95% confidence level, an intrinsic velocity dispersion of $\sigma \gtrsim 6.5 \text{ km s}^{-1}$.³ From dynamical arguments based on the virial theorem (Illingworth 1976; Richstone and Tremaine 1986, Sect. 4.1.1), such a large dispersion indicates a large dynamical mass-to-light ratio $M/L_V \gtrsim 30[M/L_V]_{\odot}$, confirming the prediction of Faber and Lin (1983) and indicating the presence of dark matter.

Figure 20-9 plots the number of stars observed in dSph stellar velocity surveys as a function of time. The three decades of observations separate neatly into “epochs” defined by the available instrumentation.

2.3.1 Small Number Statistics

The 1980s yielded the first precise velocity measurements for individual stars in Draco (Aaronson 1983), Carina, Sculptor, and Fornax (Seitzer and Frogel 1985), Leo I and Leo II (Suntzeff et al. 1986), and Ursa Minor (Aaronson and Olszewski 1987a) and follow-up observations of Sculptor (Armandroff and Da Costa 1986) and Draco (Aaronson and Olszewski 1987b). These samples typically included $\lesssim 10$ stars per galaxy and indicated velocity dispersions of $\sigma \sim 6\text{--}10 \text{ km s}^{-1}$, suggesting that dSph dynamical mass-to-light ratios reach at least double the values estimated for globular clusters. Strong implications for the particle nature of dark matter (Lin and Faber 1983; Tremaine and Gunn 1979, Sect. 6.2) drew immediate attention. Nevertheless, skepticism regarding small samples, velocity precision, and the unknown contribution of binary orbital motions to the measured velocity dispersions demanded further observations and better statistics.

³Aaronson added to the final article proof a measurement of $-285.6 \pm 1.1 \text{ km s}^{-1}$ for a fourth, non-carbon star, further supporting a large dispersion.



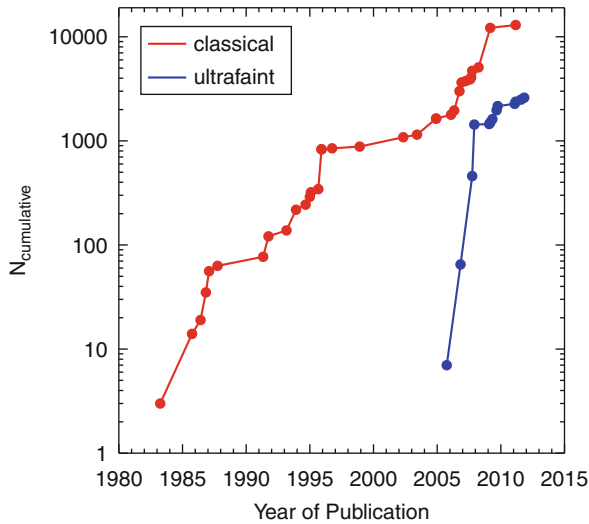
■ Fig. 20-8

Examples of spectra for individual dSph stars. *Left*: MMT echelle spectra ($R \sim 30,000$) for three carbon stars in Draco used for the first measurement of a dSph velocity dispersion (Aaronson 1983, reproduced by permission of the American Astronomical Society). *Right*: Keck/DEIMOS spectrum for the brightest red giant in Segue 1 ($R \sim 6,000$; Geha et al. 2009, reproduced by permission of the American Astronomical Society), with absorption features labeled and best-fitting model overlotted

2.3.2 Confirmation

In the 1990s, several groups accumulated velocity samples for tens of stars per dSph. Including Aaronson's original work, nine seasons of observations with the MMT eventually produced velocity samples that reached ~ 20 members in each of the Draco and Ursa Minor dSphs (Armandroff et al. 1995; Olszewski et al. 1995), yielding velocity dispersion measurements of $\sigma \sim 10 \text{ km s}^{-1}$ and dynamical mass-to-light ratios $M/L_V \sim 75[M/L_V]_{\odot}$ for both galaxies. Meanwhile, medium-resolution ($R \sim 12,000$) spectra from the William Herschel Telescope gave velocity samples for tens of stars in each of Draco, Ursa Minor, and the newly discovered Sextans (Hargreaves et al. 1994a, b, 1996b). High-resolution spectra from ESO's 3.6-m and NTT telescopes (Queloz et al. 1995) and Keck/HIRES (Mateo et al. 1998) delivered precise velocities for 23 and 33 stars in Sculptor and Leo I, respectively. In all cases, velocity dispersions of $\sigma \gtrsim 6 \text{ km ms}^{-1}$ indicated $M/L_V \gtrsim 10[M/L_V]_{\odot}$.

Providing an early demonstration of the efficiency of multi-object fiber spectroscopy, Armandroff et al. (1995) compiled samples of ~ 100 velocities in each of Draco and Ursa Minor using the HYDRA multi-fiber spectrograph at the KPNO 4-m telescope. This data set included



■ Fig. 20-9

Growth of sample sizes available for internal kinematic studies of the Milky Way's dwarf spheroidal satellites. Plotted as a function of time is the cumulative number of line-of-sight velocities measured for individual stars (including foreground contamination) targeted in dSph spectroscopic surveys (References: Aaronson (1983), Seitzer and Frogel (1985), Suntzeff et al. (1986, 1993), Armandroff and Da Costa (1986), Aaronson and Olszewski (1987a, b), Mateo et al. (1991, 1993, 1998, 2008), Da Costa et al. (1991), Hargreaves et al. (1994a, b), Armandroff et al. (1995), Vogt et al. (1995), Quéroz et al. (1995), Olszewski et al. (1995), Hargreaves et al. (1996b), Kleyna et al. (2002, 2003, 2005), Tolstoy et al. (2004), Muñoz et al. (2006a, b), Battaglia et al. (2006, 2011), Westfall et al. (2006), Walker et al. (2006, 2007a), Koch et al. (2007a, b, 2009), Simon and Geha (2007), Sohn et al. (2007), Geha et al. (2009), Walker et al. (2009a, c), Belokurov et al. (2009), Carlin et al. (2009), Willman et al. (2011), Simon et al. (2011), Adén et al. (2011), and Koposov et al. (2011b)

many repeat measurements, which Olszewski et al. (1996) used to estimate a binary fraction of ~ 0.2 – 0.3 for periods of ~ 1 year. Based on Monte Carlo simulations, Olszewski et al. (1996) and Hargreaves et al. (1996a) concluded that binary motions contribute negligibly to the velocity dispersions measured for classical dSphs (► Sect. 3.3).

In the Southern Hemisphere, Mateo et al. (1991) used Las Campanas Observatory's 2.5-m telescope and "2D Frutti" echelle photon counter to measure velocities for 44 Fornax stars, including an outer field that showed the same velocity dispersion ($\sigma \sim 10 \text{ km s}^{-1}$) as the central stars, indicating $M/L_V \sim 10[M/L_V]_{\odot}$. After measuring a velocity dispersion of $\sigma \sim 7 \text{ km s}^{-1}$ and $M/L_V \sim 40[M/L_V]_{\odot}$ from the velocities of 17 Carina stars, Mateo et al. (1993) noted a scaling relation among dSphs: the dynamical mass-to-light ratios of classical dSphs are inversely proportional to luminosity, suggesting similar dynamical masses of $\sim 10^7 M_{\odot}$ (► Sect. 5).

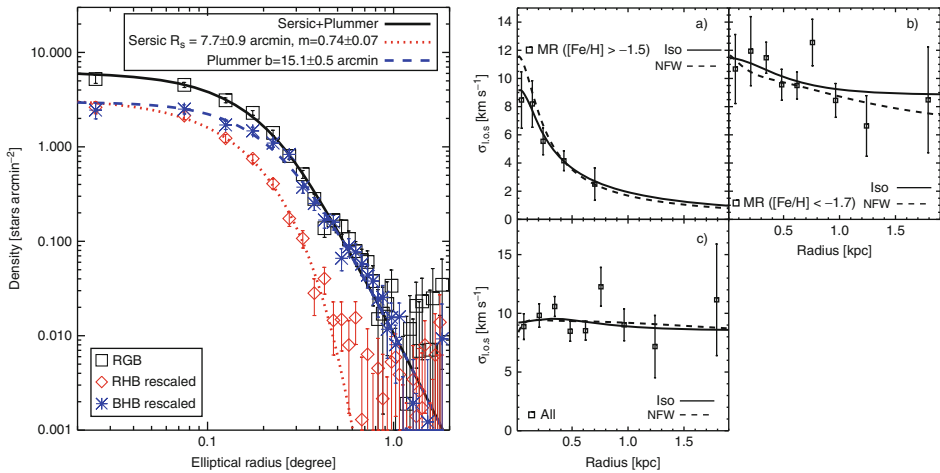
2.3.3 Large Samples

The gap in ► Fig. 20-9 between 1998 and 2002 signifies a period not of inactivity but rather of construction. During this time, wide-field multi-object spectrographs were built for the world's

largest telescopes. Over the past decade, surveys with these new instruments have increased stellar velocity samples from tens to thousands per dSph. Klynya et al. (2002, 2003, 2004) used the Wide-Field Fibre Optic Spectrograph (WYFFOS) at the 2.5-m Isaac Newton Telescope to measure velocities for ~ 100 stars in each of Draco, Ursa Minor, and Sextans, respectively. These samples were sufficiently large to examine the velocity distribution as a function of distance from the dSph center (e.g., Wilkinson et al. 2002, 2004).

Soon thereafter, the Dwarf Abundances and Radial-Velocities Team (DART) used the FLAMES fiber spectrograph at the 8.2-m Very Large Telescope (VLT; UT2) to measure velocities and metallicities (derived from the strength of the calcium-triplet absorption feature at $\sim 8,500$ Å) for ~ 310 , ~ 560 , and ~ 175 members of Sculptor, Fornax, and Sextans, respectively (Battaglia et al. 2006, 2011; Tolstoy et al. 2004). These samples yielded the discovery that all three of these dSphs contain multiple, chemodynamically independent stellar populations (► Fig. 20-10, ► Sect. 4.2.2). Also with VLT/FLAMES, Koch et al. (2007a) measured a velocity dispersion of $\sigma \sim 7$ km s $^{-1}$ from ~ 170 members of Leo II, providing what remains the largest published sample for this galaxy.



Meanwhile, Muñoz et al. (2006b) used archival VLT/FLAMES spectra (see also Fabrizio et al. 2011) to measure velocities for ~ 300 Carina members and added another ~ 45 members from spectra obtained sequentially with the MIKE spectrograph at the Magellan/Clay 6.5-m telescope. The extra members observed with MIKE extend to ~ 5 times Carina's limiting radius as determined from photometry (► Fig. 20-6), indicating that Carina is losing mass to tidal interactions with the Milky Way.



► Fig. 20-10

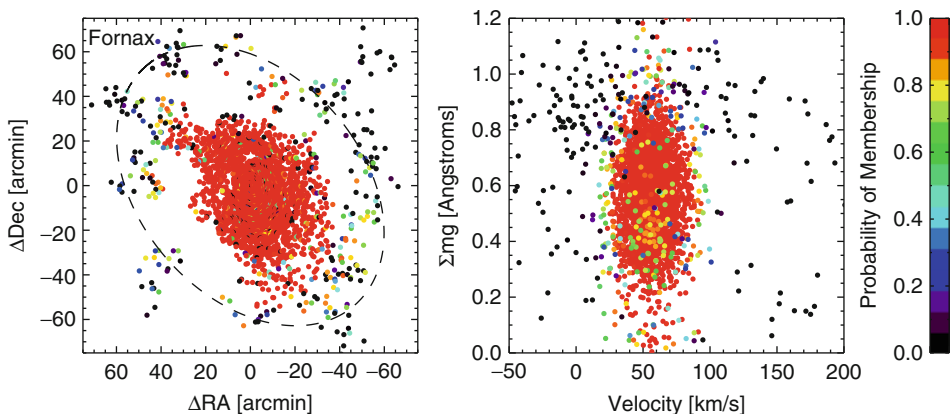
Sculptor's two chemodynamically independent stellar subpopulations (Battaglia et al. 2008, reproduced by permission of the American Astronomical Society; see also Tolstoy et al. 2004) Left: Surface brightness profiles for Sculptor's red giants (black), red-horizontal branch (red) and blue-horizontal branch stars. Right: Velocity dispersion profiles calculated for subsets of relatively metal rich (top left) and metal poor (top right) as determined from the VLT/FLAMES spectroscopic sample of Tolstoy et al. (2004). For comparison, the lower-left panel plots the velocity dispersion profile measured from the composite population


Koch et al. (2007a) used a pair of multi-slit spectrographs – the Gemini Multi-Object Spectrograph (GMOS) at the Gemini-North 8-m telescope and the Deep Imaging Multi-Object Spectrograph (DEIMOS) at the Keck 10-m telescope – to measure velocities for ~ 100 members of Leo I. Sohn et al. (2007) added another ~ 100 members from their own observations with Keck/DEIMOS, and Mateo et al. (2008) contributed velocities for ~ 300 Leo I members using the Hectochelle multi-fiber spectrograph at the MMT. The latter two studies explored larger radii, and both found kinematic evidence for tidal streaming motions in the outskirts of Leo I. This result is surprising given Leo I's current distance of ~ 250 kpc (Irwin and Hatzidimitriou 1995), leading Mateo et al. (2008) to suggest that Leo I's orbit is nearly radial.


Operating from 2004 to 2011, the Michigan-MIKE Fiber Spectrograph (MMFS; $R \sim 20,000$), built by Mario Mateo for the Magellan/Clay 6.5-m telescope, provided what remain the largest homogeneous velocity samples for “classical” dSphs. The public catalog of Magellan/MMFS velocities includes ~ 775 , $\sim 2,500$, $\sim 1,365$, and ~ 440 members of Carina, Fornax, Sculptor, and Sextans, respectively (Walker et al. 2009c).  Figure 20-11 displays the Fornax data, including sky positions as well as the two quantities measured from each spectrum: line-of-sight velocity and a spectral index that indicates the pseudo-equivalent width of the Mg-triplet feature at $\sim 5,170$ Å. Data from a similar survey conducted in the North with MMT/Hectochelle will soon become public.  Figure 20-12 displays velocity dispersion profiles calculated from the Magellan/MMFS and MMT/Hectochelle data, demonstrating that the luminous regions of classical dSphs have approximately constant velocity dispersion.

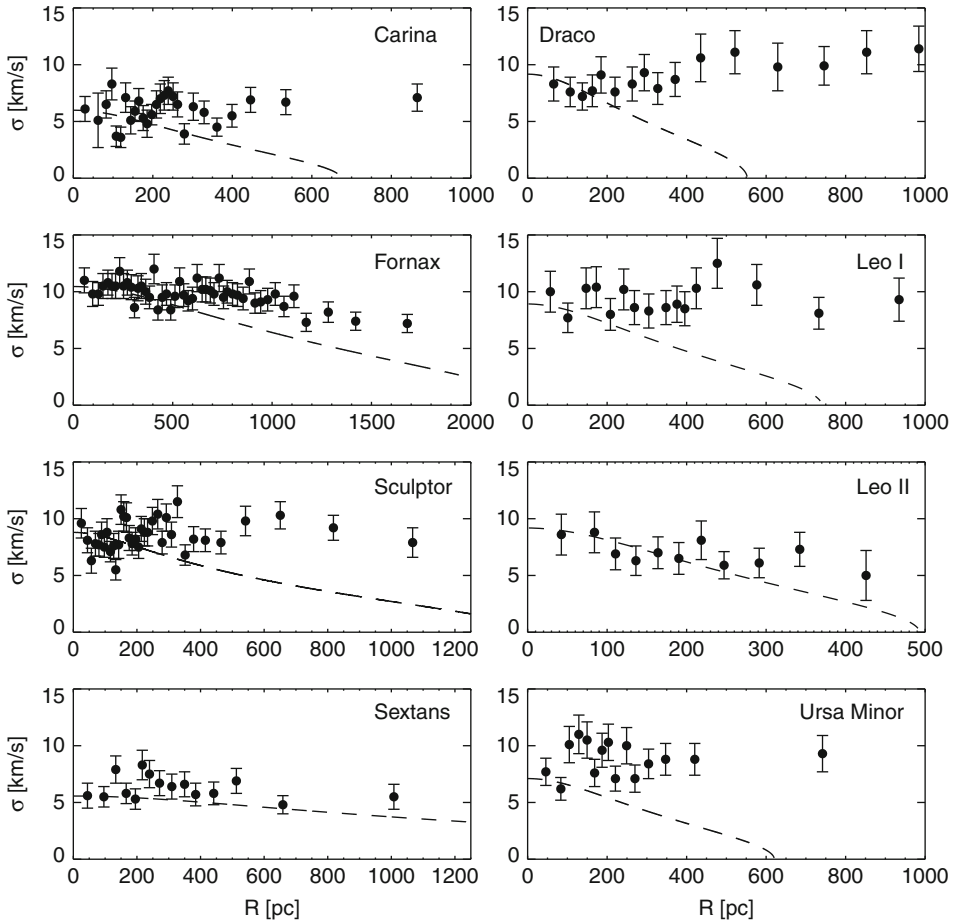
2.3.4 (Necessarily) Small Samples

With kinematic samples for “classical” dSphs growing exponentially, discoveries of “ultrafaint” dSphs with SDSS data generated a wave of interest in the faintest Milky Way satellites, and efforts



 Fig. 20-11

Magellan/MMFS spectroscopic data for Fornax (Walker et al. 2009c, updated to include the full sample of $\sim 3,200$ members). *Left*: Sky positions of individual stars. *Right*: velocities and spectral indices (pseudo-equivalent widths of the Mg-triplet absorption feature). *Color* indicates membership probability as estimated from position, velocity, and spectral index distributions. The *ellipse* indicates the limiting radius, R_K  20.1)





■ Fig. 20-12

Velocity dispersion profiles observed for the Milky Way's eight "classical" dSphs (Mateo et al. 2008; Walker et al. 2007a, 2009b). See also Kleyna et al. (2002, 2004), Wilkinson et al. (2004), Muñoz et al. (2005, 2006b), Sohn et al. (2007), Koch et al. (2007a, b), and Battaglia et al. (2008, 2011). Overplotted are mass-follows-light King (1966) models (▶ Sect. 4.1.1) normalized to reproduce the observed central dispersions. Failure of these models to reproduce the large velocity dispersions at large radius provides the strongest available evidence that dSphs have dominant and extended dark matter halos


to obtain spectroscopic follow-up began immediately. Kleyna et al. (2005) contributed a first result echoing Aaronson's original study of Draco: from Keck/HIRES velocities for five members of Ursa Major I, Kleyna et al. (2005) conclude that $\sigma > 6.5 \text{ km s}^{-1}$ with 95% confidence. Given UMaI's low luminosity, simple dynamical models imply $M/L_V \gtrsim 500 [M/L_V]_{\odot}$. Muñoz et al. (2006a) used the HYDRA multi-fiber spectrograph at the 3.5-m WIYN telescope to measure


velocities for seven members of the Boötes I dSph, obtaining a velocity dispersion of $\sim 6.5 \text{ km s}^{-1}$ and $M/L_V \gtrsim 130[M/L_V]_{\odot}$.


Whereas spectroscopic surveys of “classical” dSphs target bright red giant branch (RGB) stars, the least luminous “ultrafaint” satellites host few RGBs. Samples for even tens of stars for such objects require observations of faint stars near the main sequence turnoff and are feasible only with the largest telescopes. Martin et al. (2007) and Simon and Geha (2007) used Keck/DEIMOS to observe tens of velocities in 10 of the 11 “ultrafaints” known at the time, measuring $\sigma \gtrsim 3 \text{ km s}^{-1}$ and concluding that these objects indeed have extremely large dynamical mass-to-light ratios of order $M/L_V \gtrsim 100[M/L_V]_{\odot}$ and larger. Geha et al. (2009) and Simon et al. (2011) followed with a Keck/DEIMOS survey of Segue 1 (see example spectrum in the right-hand panel of  Fig. 20-8), concluding that this object is the “darkest galaxy,” with $M/L_V \sim 3,400[M/L_V]_{\odot}$.

Adén et al. (2009) and Koposov et al. (2011b) used VLT/FLAMES to measure velocity dispersions for Hercules and Boötes I, respectively, that while indicative of large dynamical mass-to-light ratios, were smaller than previously measured with Keck/DEIMOS. Adén et al. (2009) obtained a smaller velocity dispersion after using Strömgren photometric criteria to remove foreground interlopers. Koposov et al. (2011b) used a novel observation strategy that included ~ 15 individual 45–60 min exposures taken over a month. After measuring velocities for each exposure, Koposov et al. (2011b) were able to resolve binary orbital motions directly ( Sect. 3.3).

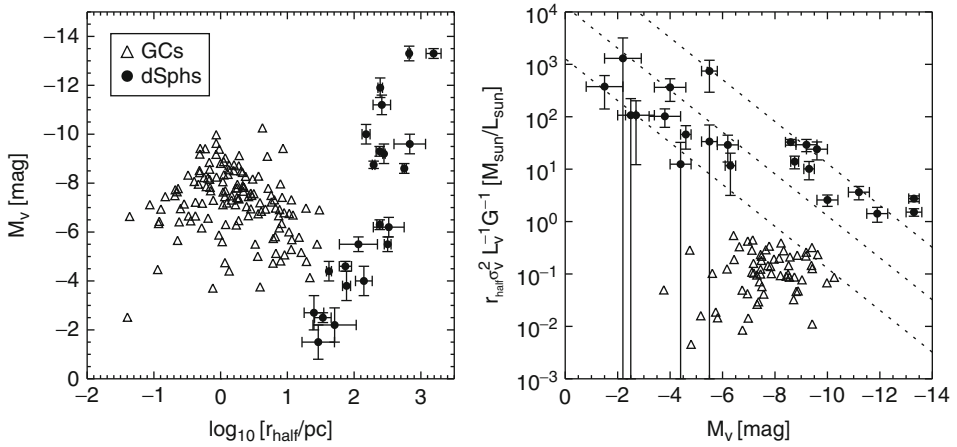
2.4 The Smallest Galaxies

 Figure 20-13 displays two scaling relations defined by the structural and kinematic observations discussed above and lets one compare the properties of objects classified as dSphs directly with those of objects classified as globular clusters. Parameters for globular clusters are adopted from the catalog of Harris (1996, 2010 edition).

The left-hand panel of  Fig. 20-13 plots luminosity against size, characterized by the projected half-light radius (Belokurov et al. 2007; Gilmore et al. 2007; Martin et al. 2008; Sand et al. 2011). Here one can appreciate another conclusion of Shapley (1938b) regarding the first known dSphs: “The Sculptor and Fornax systems might be called greatly expanded giant clusters.” Indeed, while the luminosity distributions of dSphs and globular clusters overlap substantially, most dSphs have $R_h \gtrsim 100 \text{ pc}$, while most globular clusters have $R_h \lesssim 10 \text{ pc}$.⁴ The region between $10 \lesssim R_h/\text{pc} \lesssim 100$ is populated only by globular clusters with $M_V \lesssim -4$ or by dSphs with $M_V \gtrsim -4$.

Current kinematic results suggest that this separation in luminosity between the smallest dSphs and the largest globular clusters is not merely an artifact of classification but points to a fundamental structural difference. The right-hand panel of  Fig. 20-13 plots the product $R_h \sigma^2 / (L_V G)$ (dimensionally a mass-to-light ratio) against luminosity. In the region of overlapping size, the less luminous objects tend to have larger velocity dispersions, amplifying the separation in luminosity such the smallest, faintest dSphs have the largest dynamical mass-to-light ratios of any known galaxies (Geha et al. 2009; Kleyana et al. 2005; Martin et al. 2007;

⁴M31 hosts several “extended” globular clusters with half-light radii as large as several tens of pc (e.g., Huxor et al. 2005), but no similar population within the Milky Way has yet been discovered.



■ Fig. 20-13

Left: luminosity versus size (updated from Belokurov et al. 2007; Gilmore et al. 2007; Martin et al. 2008) for Milky Way satellites including objects classified as globular clusters (*open triangles*; data from Harris (1996, 2010 edition)) and dwarf spheroidals (*filled circles with error bars*; data from compilations by Irwin and Hatzidimitriou (1995), Mateo (1998), Martin et al. (2008), and Sand et al. (2011)). The apparent lack of objects toward low luminosity and large size is a selection effect that reflects the surface-brightness limit of the SDSS survey (Koposov et al. 2008). **Right:** dynamical mass-to-light ratio (modulo of a constant scale factor) versus luminosity. Updated from Mateo et al. (1993), Mateo (1998), Simon and Geha (2007), and Geha et al. (2009). *Dotted lines* correspond to constant masses of 10^5 , 10^6 , and $10^7 M_{\odot}$

Muñoz et al. 2006a; Simon and Geha 2007; Simon et al. 2011).⁵ Furthermore, the ultrafaint dSphs extend a relation under which less luminous dSphs have larger M/L_V (Mateo 1998; Mateo et al. 1993, [Sect. 5](#)). The discontinuity in dynamical M/L_V and between dSphs and globular clusters marks a boundary between objects with dark matter and those without.

⁵Some ambiguity regarding the masses of the smallest, faintest dSphs results from the convergence of three relevant quantities – the typical velocity measurement error, the measured velocity dispersions, and the potential contribution to the measured dispersions from binary orbital motions – on the same value, $\sim 3\text{--}4 \text{ km s}^{-1}$ (Koposov et al. 2011b; Martin et al. 2007; McConnachie and Côté 2010; Simon and Geha 2007; Simon et al. 2011). For some faint dSphs, the most compelling evidence for large amounts of dark matter comes from stellar chemistry rather than kinematics. The faintest objects classified as dSphs tend to have metallicity dispersions ($\sigma_{[\text{Fe}/\text{H}]} \gtrsim 0.4 \text{ dex}$, Geha et al. 2009; Kirby et al. 2011; Norris et al. 2010; Willman et al. 2011) indicative of prolonged and perhaps multiple episodes of star formation, thereby requiring gravitational potentials sufficiently deep to retain interstellar media despite pressures generated by stellar feedback. An adequate discussion of the relationships between dSph kinematics and stellar chemistry is beyond the scope of the present work; Tolstoy et al. (2009) provide an excellent, recent review.

3 Stellar Velocity Dispersion as a Proxy for Mass

In rotating spiral galaxies, the circular velocity at radius r relates directly to enclosed mass via $v_{\text{circ}}^2 = GM(r)/r$. Ordered rotation in dSphs is dynamically negligible (see next section); instead, dSphs are supported against gravity primarily by the random motions of their stars. Therefore, the estimation of dSph masses is a fundamentally statistical enterprise. The simplest statistic that characterizes dSph stellar dynamics is the dispersion of velocities along the line of sight, σ . For a relaxed system of characteristic size R , the virial theorem implies $\sigma^2 \propto GM/R$. In principle, measurements of the size and velocity dispersion of a dSph provide a simple estimate of its mass (► Sect. 4). In practice, one must be aware of effects that can inflate measured values of the velocity dispersion above equilibrium values.

3.1 Rotation

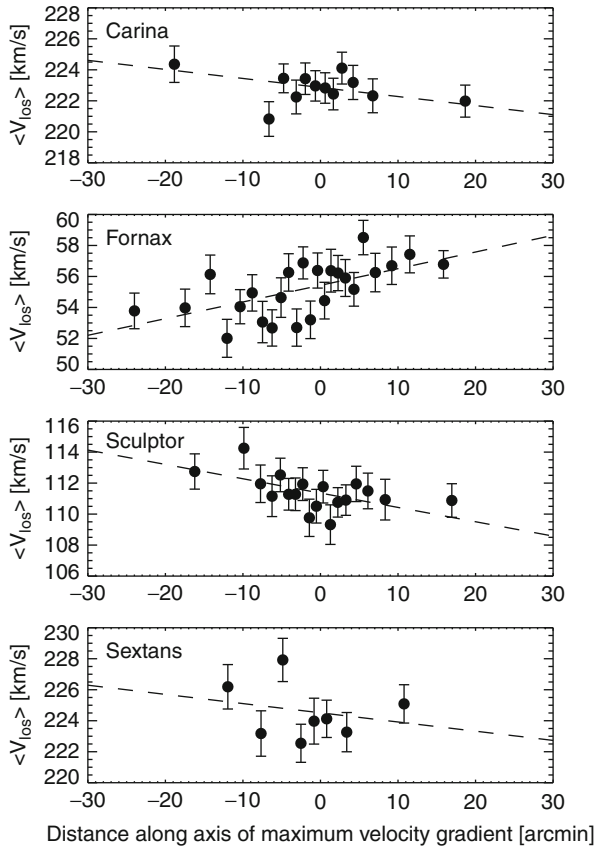
In the simplest case, solid-body rotation about an axis misaligned with the line of sight will induce a gradient, dV/dR , in the line-of-sight velocity distribution. For example, ► Fig. 20-14 plots mean velocity along the axes for which the observed velocity gradients are maximal in Carina, Fornax, Sculptor, and Sextans. Even at the radii of the outermost observed stars, any ordered motion due to rotation is limited to $R_{\text{max}}dV/dR \lesssim 3 \text{ km s}^{-1}$, negligible compared with the observed velocity dispersions of $\sigma \sim 10 \text{ km s}^{-1}$.

For Carina and Fornax, the amplitude and orientation of the observed velocity gradients are consistent with a perspective effect induced not by rotation but rather by these dSphs' systemic orbital motions transverse to the line of sight (Kaplinghat and Strigari 2008; Walker et al. 2008) as measured from HST astrometry (Piatek et al. 2002, 2003, 2007). The observed signal in Sculptor cannot be attributed to its measured proper motion (Piatek et al. 2006; Schweitzer et al. 1995), and thus Sculptor may have a residual rotational component (Battaglia et al. 2008), albeit one that contributes weakly ($v_{\text{rot}}/\sigma \lesssim 0.5$) to the measured velocity dispersion.

3.2 External Tides

All dSphs considered here orbit within the gravitational potential of the Milky Way and are therefore susceptible to external influence from tidal forces. Tides can affect the structure and kinematics of a given satellite through a variety of mechanisms, including stripping, shocking, “stirring” and various orbital resonances. Depending on the strengths of interactions and timescales of subsequent relaxation, tides might inflate observed velocity dispersions and mass estimates based thereupon.

In the most extreme scenarios, tides have been invoked to explain dSph velocity dispersions without dark matter (e.g., Fleck and Kuhn 2003; Kroupa 1997; Kuhn 1993; Kuhn and Miller 1989; Metz and Kroupa 2007). Such explanations are most plausible for dSphs that are closest to the Milky Way and exhibit elongated stellar structures, e.g., Ursa Major II (Muñoz et al. 2010; Zucker et al. 2006a), Hercules (Belokurov et al. 2007; Coleman et al. 2007; Martin and Jin 2010), and possibly Segue 1 (Niederste-Ostholt et al. 2009). However, generalisation of purely tidal mechanisms to explain the apparent dark matter content of the entire dSph population does not



■ Fig. 20-14

Line-of-sight velocity gradients in the Milky Way’s “classical” dSph satellites. Panels display mean velocity as a function of distance along the axis corresponding to the maximum velocity gradient from the samples of Walker et al. (2009c). The observed gradients give rise to maximum amplitudes $R_{\text{max}} dV/dR \lesssim 3 \text{ km s}^{-1}$ significantly less than the velocity dispersions ($\sigma \sim 10 \text{ km s}^{-1}$). Using VLT spectra from the DART survey, Battaglia et al. (2008) report similar results for Sculptor, measuring a gradient of $7.6^{+3.0}_{-2.2} \text{ km s}^{-1} \text{ deg}^{-1}$ along the morphological major axis

account without contrivance for the wide distribution of dSph distances ($\sim 30 \lesssim D/\text{kpc} \lesssim 250$) or for a monotonic metallicity-luminosity relation (Kirby et al. 2008, 2011; Mateo 1998).

While tidal stripping involves the transfer of mass from the satellite to the parent outside a particular boundary,⁶ tidal shocking involves impulsive the injection of energy into the satellite as it plunges through the disk and/or near the center of the parent system

⁶ Equation 20.4 gives the tidal radius for the idealised case of point-mass potentials. Calculations and simulations by Read et al. (2006b) demonstrate that stars are actually lost from various depths depending on the internal mass distributions of satellite and parent, as well as on the properties of the stellar orbits themselves (e.g., prograde versus retrograde with respect to the orbit of the dSph about the Galaxy).

(e.g., Gnedin et al. 1999). While the former process removes mass directly from the satellite's outer regions, the latter process tends to decrease its central density (Read et al. 2006a). However, numerical simulations by Piatek and Pryor (1995) and Oh et al. (1995) suggest that even strong tidal interactions do not significantly inflate a satellite's central velocity dispersion, which can therefore remain a reliable indicator of dynamical mass.

Many recent N-body simulations examine specific phenomenology associated with tidal interactions. For example, Read et al. (2006a) use simulations to demonstrate that the projection of tidal streaming motions along most viewing angles tends to cause velocity dispersion profiles to increase at large radius, and argue that the lack of such upturns in the classical dSphs (with the possible exception of Draco – see [Fig. 20-12](#)) limits the severity of current disruption events. Peñarrubia et al. (2008) use simulations to show that repeated tidal encounters cause monotonic declines in the satellite's central surface brightness, velocity dispersion, and scale radius (as evaluated at apocenter). Further simulations by Peñarrubia et al. (2009) indicate that when a dark matter halo is present, tides do *not* generate a clear truncation in the surface brightness profile of the bound remnant as prescribed by the dynamical models of King (1966). Rather, as the remnant relaxes after a pericentric encounter, tidal debris generates an “excess” of stars at radii where the local crossing time ($R_c/\sigma \sim 10$ Myr at the core radius of a typical dSph) exceeds the time elapsed since the encounter.

These simulations provide a context for evaluating the tidal origin of breaks and bumps in observed surface-brightness profiles of individual dSphs ([Sect. 2.2.3](#), [Fig. 20-2](#)), and for gauging the severity with which tides influence the observed kinematics. In general, one can expect tides to have stronger and more enduring influence on the outer regions of dSphs, and for the degree of influence on any particular satellite to scale with orbital parameters. A typical dSph, with $R_c \sim 100$ pc and $\sigma \sim 10$ km s⁻¹, requires ~ 20 core crossing times to travel a distance of ~ 100 kpc (a typical Galactocentric distance) at speed ~ 250 km s⁻¹; the assumption of dynamic equilibrium should therefore hold reasonably well out to several core radii for the majority of dSphs. In practice one must use all available information about a satellite's orbit and outer structure to evaluate the likely contribution of tides on a case-by-case basis.

3.3 Binary Stars

Unresolved binary orbital motions might contribute significantly to the observed velocity dispersions of the lowest-mass galaxies. Olszewski et al. (1995) use 112 independent velocity measurements for 42 stars in Draco and Ursa Minor to identify seven stars that exhibited velocity variability. Elimination of these stars from their samples has negligible impact on the measured velocity dispersions of Draco and Ursa Minor. Using the larger KPNO/HYDRA sample of Armandroff et al. (1995) (373 independent velocity measurements for 185 stars), Olszewski et al. (1996) perform Monte Carlo simulations to estimate that the binary frequency for Draco and Ursa Minor stars with periods of ~ 1 year is ~ 0.2 – 0.3 per decade of period. Even though this fraction is larger than the one found in the solar neighborhood (e.g., Duquenoey and Mayor 1991), further simulations by Olszewski et al. (1996) and Hargreaves et al. (1996a) demonstrate that the scatter introduced by binaries is small compared to the measured dispersions of $\sigma \sim 10$ km s⁻¹ and thus that binaries do not significantly inflate dynamical masses of the “classical” dSphs.

More recently, measurements of velocity dispersions as small as $\sigma \sim 3$ km s⁻¹ for several ultrafaint satellites (Adén et al. 2009; Koposov et al. 2011b; Martin et al. 2007; Simon and Geha 2007; Walker et al. 2009a) have renewed concerns over the possible contribution of binary

motions. While simulations by Minor et al. (2010) find that binaries have little effect on dispersions measured for systems with intrinsic velocity dispersions $\sigma \gtrsim 4 \text{ km s}^{-1}$, complementary simulations by McConnachie and Côté (2010) demonstrate that for systems with intrinsic dispersions near zero (as would be expected for many of the ultrafaints if they contain no dark matter), binaries can inflate measured dispersions to values as high as $\sigma \sim 4 \text{ km s}^{-1}$.

It is therefore necessary to verify the extreme mass-to-light ratios of the faintest (and coldest) dSphs with repeat spectroscopic measurements that constrain the velocity variability of individual stars. Simon et al. (2011) present second-epoch Keck/DEIMOS velocity measurements for several of the Segue 1 stars first measured by Geha et al. (2009). They find that one of Segue 1's six red giants shows significant velocity variability, as do two fainter stars. Including parametric binary orbital distribution functions in a Bayesian analysis of Segue 1's velocity dispersion, Simon et al. (2011, see also Martinez et al. 2011) estimate a velocity dispersion of $3.7^{+1.4}_{-1.1} \text{ km s}^{-1}$, implying a dynamical mass-to-light ratio of $M/L_V \sim 3,400[M/L_V]_{\odot}$ and reinforcing the notion that Segue 1 is the “darkest” galaxy known.

In a separate study, Koposov et al. (2011b) use ~ 15 VLT/FLAMES observations obtained over 1 month in order to resolve binary motions directly among members of Boötes I. **Figure 20-15** displays independent velocity measurements for two stars as a function of time. While velocities for the star in the left-hand panel are consistent with a constant velocity, velocities for the star in the right-hand panel change systematically by $\sim 10 \text{ km s}^{-1}$. After discarding probable binaries, Koposov et al. (2011b) find that most members of Boötes I belong to a cold population with dispersion $\sigma = 2.4^{+0.9}_{-0.5} \text{ km s}^{-1}$ (**Fig. 20-16**), significantly smaller than previous single-epoch estimates of $\sigma \sim 6.5 \text{ km s}^{-1}$ (Martin et al. 2007; Muñoz et al. 2006a). At present, the contribution of binaries to the velocity dispersions measured for the coldest dSphs is poorly understood, and more multi-epoch studies are required.

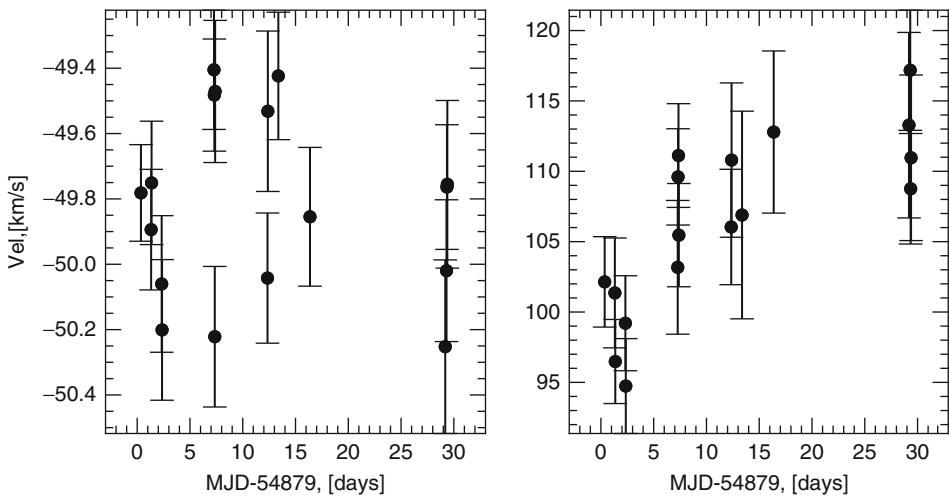
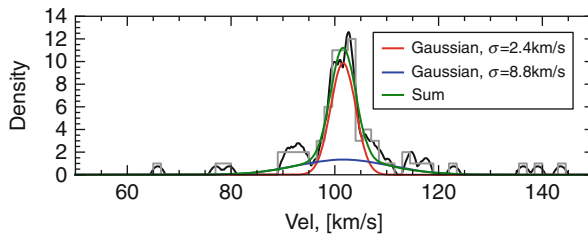


Fig. 20-15

Direct detection of binary stars from individual velocity measurements in Boötes I (Koposov et al. 2011b, reproduced by permission of the American Astronomical Society). *Left*: foreground star with velocities consistent with a constant value. *Right*: member star ($\langle v \rangle_{\text{Boo}} \sim 105 \text{ km s}^{-1}$) exhibiting a velocity increase of $\sim 10 \text{ km s}^{-1}$ over 1 month (See **Sect. 3.3** for a discussion of the effects of binary orbital motions on measurements of dSph velocity dispersions)



■ Fig. 20-16

Velocity distribution of the Boötes I dSph from the VLT/FLAMES data of Koposov et al. (2011b, reproduced by permission of the American Astronomical Society), not including 11 stars observed to have significant velocity variability. In the best-fitting two-Gaussian model, 70% of the stars belong to a “cold” component with $\sigma \sim 2 \text{ km s}^{-1}$ and the rest belong to a hotter component with $\sigma \sim 9 \text{ km s}^{-1}$

4 dSph Masses

For a collisionless stellar system in dynamic equilibrium, the gravitational potential, Φ , relates to the phase-space distribution of stellar tracers,⁷ $f(\vec{r}, \vec{v}, t)$, via the collisionless Boltzmann equation (Eq. 4–13c of Binney and Tremaine 2008):

$$\frac{\partial f}{\partial t} + \vec{v} \cdot \nabla f - \nabla \Phi \cdot \frac{\partial f}{\partial \vec{v}} = 0. \quad (20.5)$$

Current instrumentation resolves the internal distributions of neither distance nor proper motions for dSph stars. The structural and kinematic observations described in Sect. 2 provide information only about the projections of phase-space distributions along lines of sight, limiting knowledge about f and hence also about Φ . Therefore, all efforts to translate existing data sets into constraints on Φ involve simplifying assumptions. Along with dynamic equilibrium, common assumptions include spherical symmetry and particular functional forms for the distribution function and/or the gravitational potential. The most useful analyses identify the least restrictive assumptions that are appropriate for a given data set. Modern structural/kinematic data contain information that is sufficient to place reasonably robust constraints not only on the amount of dSph mass but in some cases also on its spatial distribution.

4.1 Amount

4.1.1 “Mass-Follows-Light” Models

A common method for analysing dSph kinematics employs the following assumptions:

1. Dynamic equilibrium
2. Spherical symmetry
3. Isotropy of the velocity distribution, such that $\langle v_r^2 \rangle = \langle v_\theta^2 \rangle = \langle v_\phi^2 \rangle$

⁷The distribution function is defined such that $f(\vec{r}, \vec{v}, t) d^3 \vec{x} d^3 \vec{v}$ specifies the number of stars inside the volume of phase-space $d^3 \vec{x} d^3 \vec{v}$ centered on (\vec{x}, \vec{v}) at time t .

4. A single stellar component
5. The mass density profile, $\rho(r)$, is proportional to the luminous density profile, $\nu(r)$ (i.e., M/L is constant or “mass follows light”)

Historically, assumption (5) has been adopted when the velocity dispersion profile $\sigma(R)$ is unavailable. Examples include early analyses of classical dSphs (Mateo 1998 and references therein) and initial analyses of ultrafaint dSphs (e.g., Kleyna et al. 2005; Martin et al. 2007; Muñoz et al. 2006a; Simon and Geha 2007). Under this assumption, the steeply falling outer-surface-brightness profiles of dSphs (► Sect. 2.2.1, ► Fig. 20-2) motivate the use of dynamical models that allow for truncation by external tides (e.g., King 1966; Michie 1963). Consider, for example, the model of King (1966, see also Chapter 4 of Binney and Tremaine 2008), in which the distribution function depends only on energy:

$$f(\varepsilon) = k(2\pi v_s^2)^{-3/2}(\exp[\varepsilon/v_s^2] - 1). \quad (20.6)$$

The relative potential, $\Psi \equiv \Phi_0 - \Phi$, is defined such that $\varepsilon \equiv \Psi - \frac{1}{2}v^2 \geq 0$ at radii $r \leq R_K$. Given the assumption that mass follows light, this model is fully specified by dimensionless parameter $\Psi(0)/v_s^2$ (or equivalently, R_K/R_c), the core radius and one of either the central velocity dispersion or central mass density, which are all related via $R_c^2 = \frac{9\sigma_0^2}{4\pi G \rho_0}$.

Illingworth (1976) shows that under assumptions (1)–(5) and (► 20.6), the total mass is given by

$$\frac{M_{\text{tot}}}{M_\odot} = 167\eta \left[\frac{R_c}{\text{pc}} \right] \left[\frac{\sigma^2}{\text{km}^2 \text{s}^{-2}} \right]. \quad (20.7)$$

The parameter η is determined by the concentration $c \equiv \log_{10}[R_K/R_c]$. Following Mateo (1998), many authors adopt $\eta \approx 8$, which is appropriate for the low concentrations characteristic of classical dSphs but is generally not well constrained for the faintest dSphs.

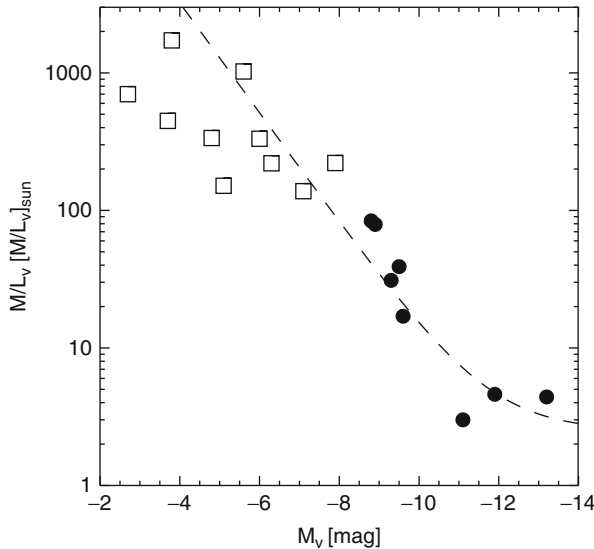
More generally, Richstone and Tremaine (1986) show that for “almost any” spherical, isotropic system with constant mass-to-light ratio and centrally cored luminosity profile, the mass-to-light ratio is approximately

$$\frac{M}{L} \approx \frac{9\sigma_0^2}{2\pi G \Sigma(0) R_{\text{hb}}}, \quad (20.8)$$

where the half-brightness radius is defined by $\Sigma(R_{\text{hb}}) = \frac{1}{2}\Sigma(0)$ and is often similar (within ~25%) to R_c . ► Figure 20-17 plots dynamical mass-to-light ratios calculated from (► 20.8) against dSph luminosity. Masses obtained for the Milky Way’s eight classical dSphs are all $M_{\text{tot}} \sim 10^7 M_\odot$. For the ultrafaints, masses range from $10^5 \lesssim M_{\text{tot}}/M_\odot \lesssim 10^7$. Dynamical mass-to-light ratios increase monotonically with decreasing luminosity, ranging from $10 \lesssim M/L_V/[M/L_V]_\odot \lesssim 1,000$.

4.1.2 Does Mass Follow Light?

When available, empirical velocity dispersion profiles provide simple tests of the assumptions listed above. Dashed lines in ► Fig. 20-12 display best-fitting King (1966, (► 20.6)) models constructed under assumptions (1)–(5), fit to the surface-brightness profiles of Irwin and Hatzidimitriou (1995, ► Fig. 20-2) and normalized to fit the central velocity dispersions. For



■ Fig. 20-17

Dynamical mass-to-light ratio derived from mass-follows-light models (● Sect. 4.1.1) versus luminosity. Data for the Milky Way's "classical" dSphs (filled circles) are from the review of Mateo (1998). Data for the 'ultrafaint' dSphs (open squares) are from Keck/DEIMOS observations by Simon and Geha (2007) and Martin et al. (2007). The dotted line corresponds to $[M/L_V]/[M/L_V]_{\odot} = 2.5 + 10^7/(L/L_{V,\odot})$ (● Sect. 5)

all of the Milky Way's classical dSphs, the central velocity dispersions imply large mass-to-light ratios $M/L_V \gtrsim 10[M/L_V]_{\odot}$ as plotted in ● Fig. 20-17. However, the mass-follows-light models underpredict the observed velocity dispersions at large radii, which tend to remain approximately constant out to the last measured points.

These discrepancies between empirical velocity dispersion profiles and mass-follows-light models in ● Fig. 20-12 imply that at least one of assumptions (1)–(5) is invalid. In fact, *all* are invalid at some level. However, various studies indicate that it is unlikely any one of assumptions (1)–(4) alone can be the problem. For example:

1. Simulations by Read et al. (2006a) suggest that tidal disruption (a violation of the equilibrium assumption) is more likely to generate rising velocity dispersion profiles rather than the flat profiles observed for real dSphs.
2. Axisymmetric models of Fornax considered by Jardel and Gebhardt (2011) favor dark matter halos that, while violating the assumption of spherical symmetry, extend well beyond the luminous component and therefore also invalidate the mass-follows-light assumption.
3. Evans et al. (2009) derive analytically an expression for the anisotropy profile, $\beta_a(r) \equiv 1 - \langle v_{\theta}^2 \rangle / \langle v_r^2 \rangle$ in terms of surface brightness and velocity dispersion profiles. If mass follows light, then the flat empirical velocity dispersion profiles tend to imply unphysical anisotropies of $\beta_a > 1$.
4. While recent observations indicate that some dSphs contain at least two distinct stellar subpopulations (● Sects. 2.2.4 and ● Sect. 4.2.2.), scale radii of the individual subpopulations

are sufficiently well constrained (e.g., Battaglia et al. 2006, 2008) that superpositions of two mass-follows-light models continue to underpredict the observed velocity dispersions at large radii. Indeed, Battaglia et al. (2008) find that the flat velocity dispersion profile they measure for Sculptor’s more spatially extended subpopulation continues to imply an even more extended dark matter halo.

Models that allow for sufficiently extended dark matter halos (violating assumption (5) while retaining assumptions (1)–(4)) can provide good fits to surface brightness and velocity dispersion profiles simultaneously (e.g., Pryor and Kormendy 1990; Wilkinson et al. 2002). On these grounds, the empirical velocity dispersion profiles shown in [Fig. 20-12](#) provide the strongest available evidence that dSphs have dominant dark matter halos that extend beyond luminous regions.

Some scenarios for dSph formation and evolution, particularly the tidal stirring mechanism of Mayer et al. (2001a, b, [Sect. 3.2](#)) and the tidal disruption simulations of Muñoz et al. (2008), tend to produce configurations in which mass approximately follows light. The results discussed above would seem to rule out this configuration. However, Łokas (2009) finds reasonable agreement with mass-follows-light models in Carina, Fornax, Sculptor, and Sextans after trimming velocity samples in order to remove member stars classified by an iterative mass estimator (Klimontowski et al. 2007) as tidally unbound. This result rests in part on a circular argument, as the adopted mass estimator (Heisler et al. 1985) is based on the virial theorem, which itself assumes mass follows light. However, the same charge of circularity can be brought against the standard kinematic analysis, in which the inclusion of stars at large radius in the kinematic analysis implicitly assumes they are bound by a sufficiently extended dark matter halo. Thus conclusions regarding the extended structure of dSph dark matter halos are generally sensitive to the assumptions employed when determining which stars to consider or reject in kinematic analyses. More secure is the conclusion that dark matter dominates dSph potentials: even mass-follows-light models require central mass-to-light ratios $M/L_V \gtrsim 10[M/L_V]_{\odot}$ in order to fit the central velocity dispersions of dSphs (Łokas 2009; Muñoz et al. 2008, [Fig. 20-12](#)).

4.1.3 Jeans Analysis

The methods for mass estimation described in the previous section either employ directly or are derived from specific distribution functions $f(\vec{r}, \vec{v})$ that correspond to physical dynamical models restricted by particular assumptions. Integration of [\(20.5\)](#) over velocity space provides an alternative starting point in the form of the Jeans equations (see Binney and Tremaine 2008). With spherical symmetry, one obtains

$$\frac{1}{v} \frac{d}{dr} (v \langle v_r^2 \rangle) + 2 \frac{\beta_a \langle v_r^2 \rangle}{r} = - \frac{GM(r)}{r^2}, \quad (20.9)$$

where $v(r)$, $\langle v_r^2 \rangle(r)$, and $\beta_a(r) \equiv 1 - \langle v_{\theta}^2 \rangle / \langle v_r^2 \rangle$ describe the three-dimensional density, radial velocity dispersion, and orbital anisotropy, respectively, of the (stellar) tracer component. The mass profile, $M(r)$, includes contributions from any dark matter halo. While there is no requirement that mass follow light, there is also no guarantee that a given solution to [\(20.9\)](#) – even one that fits the data – corresponds to a physical dynamical model (i.e., one for which $f(\vec{r}, \vec{v})$ is nonnegative).

► Equation 20.9 has general solution (Mamon and Łokas 2005; van der Marel 1994)

$$v\langle v_r^2 \rangle = \frac{1}{f(r)} \int_r^\infty f(s)v(s) \frac{GM(s)}{s^2} ds, \quad (20.10)$$

where $f(r) = 2f(r_1) \exp \int_{r_1}^r \beta_a(s) s^{-1} ds$. Projecting along the line of sight, the mass profile relates to observable profiles, the projected stellar density, $\Sigma(R)$ (► Fig. 20-2), and velocity dispersion, $\sigma(R)$ (► Fig. 20-12), according to (Binney and Tremaine 2008)

$$\sigma^2(R)\Sigma(R) = 2 \int_R^\infty \left(1 - \beta_a \frac{R^2}{r^2}\right) \frac{v\langle v_r^2 \rangle r}{\sqrt{r^2 - R^2}} dr. \quad (20.11)$$

► Equation 20.11 forms the basis for many methods of mass estimation, including parametric (e.g., Battaglia et al. 2008; Koch et al. 2007a, b; Martinez et al. 2011; Strigari 2010; Strigari et al. 2006, 2008a; Walker et al. 2007b, 2009b) and nonparametric (e.g., Wang et al. (2005)) techniques as well as algebraic inversion (e.g., Gilmore et al. 2007; Wilkinson et al. 2004).

All methods based on (► 20.11) are limited fundamentally by a degeneracy between the function of interest, $M(r)$, and the anisotropy profile, $\beta_a(r)$, which is poorly constrained by velocity data confined to the line of sight.⁸ Consideration of a common parametric method helps to illustrate this limitation. For example, it is common to assume that the the gravitational potential is dominated everywhere by a dark matter halo with mass density profile

$$\rho(r) = \rho_s \left(\frac{r}{r_s}\right)^{-\gamma} \left[1 + \left(\frac{r}{r_s}\right)^\alpha\right]^{\frac{\gamma-\beta}{\alpha}}, \quad (20.12)$$

i.e., the generalisation by Zhao (1996) of the Hernquist (1990) profile. ► Equation 20.12 provides a flexible halo model in the form of a split power law, with free parameter α controlling the transition from index $-\gamma$ at small radii ($r \ll r_s$) to a value of $-\beta$ at large radii ($r \gg r_s$).

From spherical symmetry, the density profile specifies the mass profile via

$$M(r) = 4\pi \int_0^r s^2 \rho(s) ds \quad (20.13)$$

and the surface brightness profile specifies the (deprojected) stellar density profile via

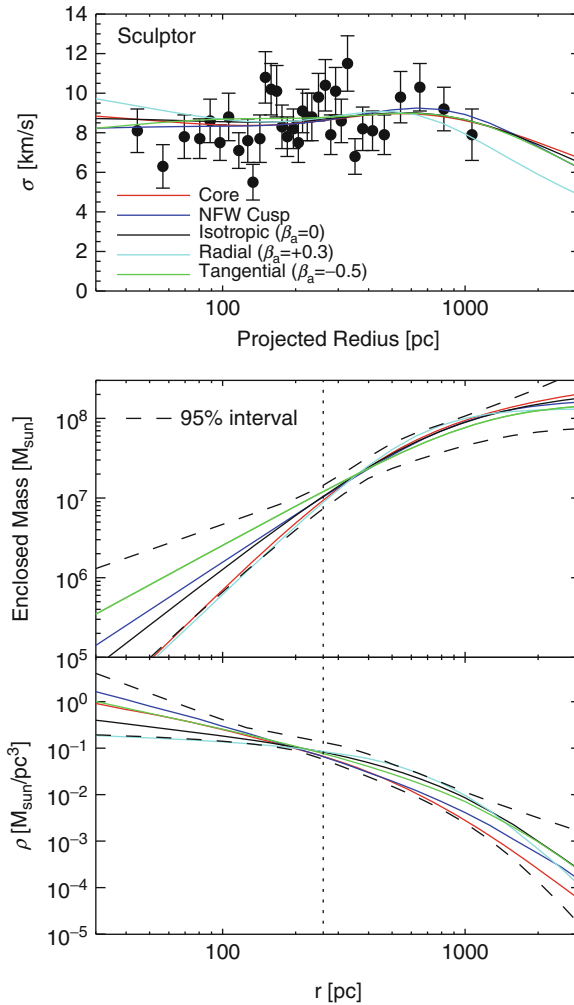
$$v(r) = -\frac{1}{\pi} \int_r^\infty \frac{d\Sigma}{dR} \frac{dR}{\sqrt{R^2 - r^2}}. \quad (20.14)$$

Given values for free parameters ρ_s , r_s , α , β , and γ and an assumption about the otherwise unconstrained anisotropy profile,⁹ (► 20.11) specifies a model velocity dispersion profile that can then be compared with observations.

Using Sculptor as an example, ► Fig. 20-18 demonstrates the degeneracy that is inherent in the standard Jeans analysis. Overplotted on Sculptor's empirical velocity dispersion profile are best-fitting models obtained under specific assumptions either about the inner slope of the mass-density profile ($\gamma = 0$ or $\gamma = 1$) or about the amount of velocity anisotropy ($\beta_a = -0.5$, $\beta_a = 0$, or $\beta_a = +0.3$). Despite corresponding to a wide range of total masses and mass distributions (bottom panels of ► Fig. 20-18), all of these models can provide equivalent fits to the structural and kinematic data.

⁸Łokas et al. (2005) develop a Jeans analysis that uses higher-order velocity moments (e.g., $\langle v^4 \rangle$) in order to reduce degeneracy between anisotropy (assumed to be constant) and total mass (effectively normalizing a cusped mass profile assumed to have $\gamma = 1$ in the notation of Eq. 20.12).

⁹Typical assumptions about anisotropy range in simplicity from $\beta_a = 0$ or $\beta_a = \text{constant}$ to $\beta_a(r) = (\beta_\infty - \beta_0)r^2/(r_\beta^2 + r^2) + \beta_0$ (e.g., Strigari 2010), introducing as many as three new free parameters.



■ Fig. 20-18

Degeneracy in the Jeans analysis of dSph mass profiles (► Sect. 4.1.3). In the *top panel*, overplotted on Sculptor's empirical velocity dispersion profile are best-fitting cored ($\gamma = 0$ in the notation of ► 20.12) and cusped ($\gamma = 1$) density profiles, as well as models that assume isotropic ($\beta_a = 0$), radially anisotropic ($\beta_a = +0.3$), and tangentially anisotropic ($\beta_a = -0.5$) velocity distributions. All of these models provide equivalent fits to the data. The *bottom two panels* plot mass and density profiles corresponding to each model. *Dotted curves* indicate 95% confidence intervals as determined from a Markov-Chain Monte Carlo analysis that assumes constant anisotropy and lets all five halo parameters (► 20.12) vary freely. The *vertical dotted line* indicates Sculptor's projected half-light radius

Notice that while the Jeans analysis provides useful constraints on none of the halo parameters included in (20.12), it does imply a model-independent constraint on one quantity. All models that fit the velocity dispersion profile have approximately the same value for the enclosed mass near the half-light radius (middle panel of Fig. 20-18), indicating that this quantity is well determined by the available data. Indeed, formal confidence intervals derived from Markov-Chain Monte Carlo scans of the full parameter space (e.g., Martinez et al. 2011; Strigari 2010; Strigari et al. 2007a; Walker et al. 2009b; Wolf et al. 2010) show a characteristic “pinch” near the half-light radius (Fig. 20-18).

Equivalent constraints on the enclosed mass at the half-light radius can be obtained more simply by solving (20.9) under the assumptions that $\beta_a = 0$ and $\sigma(R) = \text{constant}$. For a Plummer surface brightness profile (20.3), one obtains (Walker et al. 2009b)

$$M(R_h) = 5R_h\sigma^2/(2G). \quad (20.15)$$

Wolf et al. (2010) show analytically that for various surface brightness profiles, the tightest constraint on the mass profile (provided that the velocity dispersion profile is sufficiently flat) can be approximated by

$$M(r_3) = \frac{3r_3\sigma^2}{G}, \quad (20.16)$$

where r_3 is the radius at which $d \ln v / d \ln r = -3$. For most commonly adopted surface brightness profiles, this radius is close to the *deprojected* half-light radius (i.e., the radius of the sphere containing half of the stars), which typically exceeds the projected half-light radius by a factor of $\sim 4/3$. Insofar as the assumption of flat velocity dispersion profiles (and the usual assumptions of dynamic equilibrium and spherical symmetry) holds, these simple mass estimators can be applied even to the relatively sparse kinematic data available for ultrafaint satellites. Walker et al. (2009b, see erratum for updated values) and Wolf et al. (2010) tabulate masses obtained from (20.15) and (20.16), respectively, for ~ 2 dozen Local Group dSphs.

The equivalence of such seemingly crude estimates to constraints from Jeans/MCMC explorations of parameter space follows from a combination of facts: (1) Equation 20.9 deals only with velocity moments of the phase-space distribution function and not with the distribution function itself; (2) we lack empirical information about β_a ; (3) therefore, the flat velocity dispersion profiles of dSphs effectively reduce the available kinematic information to just two numbers, σ and the scale radius that characterizes the adopted surface brightness profile. The information extracted from the Jeans analysis naturally amounts to a simple combination of these two numbers.

4.2 Distribution

Some cosmological and particle physics models make specific predictions about how dark matter is distributed within individual halos (Sect. 6). Insofar as dSphs represent the structures most dominated by dark matter and least affected by the presence of baryons, their internal dynamical properties provide the most straightforward tests of such predictions. Section 4.1.3 demonstrates that so long as dSphs have flat velocity dispersion profiles, the standard Jeans analysis constrains only one number, characterizing the amount but not the distribution of dark matter. However, the restrictive assumptions employed in the standard Jeans analysis overlook structure that is present in the data available for many dSphs (Sect. 2.2.4). Recent analyses that are devised to exploit more fully the available empirical information have been able

to place useful constraints not only on the amount but also on the distribution of dark matter within individual dSphs.

4.2.1 Indirect Constraints

The presence of stellar substructure within dSphs can provide extra leverage not easily exploited in the context of equilibrium dynamical models. For example, it has long been known that Ursa Minor’s stellar component is “lumpy” (Hodge 1964b; Irwin and Hatzidimitriou 1995; Olszewski and Aaronson 1985; Palma et al. 2003, [▶ Sect. 2.2.4](#)). While Ursa Minor has a velocity dispersion of $\sigma \sim 10 \text{ km s}^{-1}$, Kleyana et al. (2003) find that a secondary peak in Ursa Minor’s stellar density field, offset by $\sim 20'$ ($\sim 400 \text{ pc}$) from the nominal center, exhibits a cold dispersion of $\lesssim 1 \text{ km s}^{-1}$. Interpreting this feature as a loosely bound star cluster captured by Ursa Minor, Kleyana et al. (2003) perform N-body simulations to examine how its stability depends on the external gravitational potential (assumed to be dominated by dark matter) of Ursa Minor. Whereas simulated clusters remain intact for a Hubble time when the host potential has a central “core” with constant density ($\gamma = 0$ in the notation of ([▶ 20.12](#))) on scales larger than the orbital radius, they disrupt on timescales of $\lesssim 1 \text{ Gyr}$ when the host potential has a centrally divergent or “cusped” density profile ($\gamma > \frac{1}{2}$). Therefore, if the observed clump indeed corresponds to a star cluster free of its own dark matter component, then its survival provides indirect evidence that the dark matter density of Ursa Minor is constant over the central few-hundred pc.

Substructure in the form of Fornax’s five globular clusters provides another example of indirect evidence for a cored dSph potential. Four of the clusters are projected near (within a factor of ~ 2) Fornax’s half-light radius ($R_h \sim 670 \text{ pc}$; Irwin and Hatzidimitriou 1995). Hernandez and Gilmore (1998) show analytically that the rate at which the orbits of such clusters decay due to dynamical friction depends on the underlying dSph potential. Subsequent numerical simulations by Sánchez-Salcedo et al. (2006) and Goerdt et al. (2006) demonstrate that in a cusped Fornax potential, dynamical friction would require only a few Gyr to bring the clusters from their present positions (assuming the projected distances from Fornax’s center are not much smaller than the true distances) all the way to Fornax’s center. On the other hand, in a cored potential, dynamical friction would bring the clusters only as close as the core radius where the harmonic potential inhibits further decay. It should be noted that further simulations by Goerdt et al. (2010), and Cole et al. (2011) suggest that the transfer of angular momentum from a sinking cluster to the central dark matter is capable of *transforming* an originally cusped into a cored potential, a possibility that has consequences for the interpretation of cored dark matter halos in dSphs ([▶ Sect. 6.1](#)).

4.2.2 Constraints from Models

Several groups have recently developed dynamical and/or kinematic models that exploit particular structure that is present in dSph data but not considered in the Jeans analyses discussed in [▶ Sect. 4.1.3](#). For example, the discoveries of two chemodynamically independent stellar subpopulations ([▶ Sect. 2.2.4](#)) in Sculptor (Tolstoy et al. 2004), Fornax (Battaglia et al. 2006), and Sextans (Battaglia et al. 2011) enable analyses of two tracer components in the same potential. After imposing a metallicity cutoff to separate Sculptor’s two subpopulations, Battaglia et al. (2008) find that cored rather than cusped potentials provide better simultaneous fits in a Jeans analysis ([▶ Sect. 4.1.3](#)) of the two sets of surface brightness and velocity dispersion profiles.

Using the same empirical profiles for Sculptor, Amorisco and Evans (2012) confirm this result by modeling both subpopulations with anisotropic King-Michie (King 1962, 1966; Michie 1963) distribution functions. These dynamical models again favor cored ($\gamma = 0$) rather than cusped ($\gamma = 1$) potentials, with a likelihood ratio sufficient to reject the hypothesis of cusped potentials with confidence $\gtrsim 99\%$.

Jardel and Gebhardt (2011) take a different approach, constructing axisymmetric three-integral Schwarzschild (1979) models for both cored and cusped potentials that also allow for a central black hole. For a given potential, libraries of stellar orbits are calculated and each orbit receives a weight based on fits to the observed distribution of velocities within discretely binned radii. Notice that while the Jeans analysis is sensitive only to the variance of the velocity distribution in a given bin, the Schwarzschild method is sensitive to the *shape* of the distribution.¹⁰ Jardel and Gebhardt (2011) find that their models constructed from cored potentials fit Fornax's velocity data significantly better than those constructed from cusped potentials. Within the context of their adopted models, they also place an upper limit of $\lesssim 3.2 \times 10^4 M_\odot$ on the mass of any central black hole.

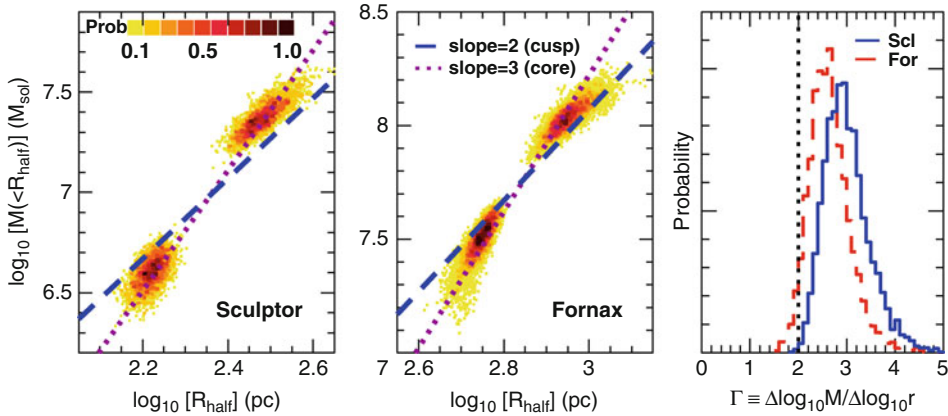
4.2.3 Direct Measurement

Walker and Peñarrubia (2011) introduce a method for measuring the slopes of dSph mass profiles that operates directly on discrete measurements of stellar positions, velocities, and spectral indices and does not invoke a dark matter halo model. This method works by combining two previous results: (1) the product of half-light radius and (squared) velocity dispersion gives a robust estimate of the mass enclosed within the half-light radius (☛ Sect. 4.1.3) and (2) some dSphs contain two independent stellar subpopulations characterized by different half-light radii, velocity dispersions, and metallicities (☛ Sect. 2.2.4). Walker and Peñarrubia (2011) use measurements of stellar positions, velocities and spectral indices to distinguish subpopulations in the Fornax and Sculptor dSphs. Estimates of half-light radii and velocity dispersions for two subpopulations provide mass estimates $M(R_h) \propto R_h \sigma^2$ at two different radii in the same mass profile immediately defining a slope. For Fornax and Sculptor, this method yields slopes of $\Gamma \equiv \Delta \log M / \Delta \log r = 2.61_{-0.37}^{+0.43}$ and $\Gamma = 2.95_{-0.39}^{+0.51}$, respectively, on scales defined by values $\sim 0.2 \lesssim R_h/\text{kpc} \lesssim 1$ estimated for the half-light radii. These slopes are consistent with cored ($\gamma = 0$) potentials, for which $\Gamma \leq 3$ at all radii, but incompatible with cusped ($\gamma \gtrsim 1$) potentials, for which $\Gamma \leq 2$ (☛ Fig. 20-19).

5 A Common dSph Mass?

Mateo et al. (1993) noticed that dSph dynamical mass-to-light ratios increase monotonically with decreasing luminosity (left panel of ☛ Fig. 20-17). Based on mass-follows-light analyses (☛ Sect. 4.1.1), the Milky Way's eight classical dSphs trace a relationship $\log_{10}(M/L_V/[M/L_V]_\odot) \sim 2.5 + 10^7 L_V/L_{V,\odot}$ (Mateo 1998). This relation implies that if their stellar mass-to-light ratios are $M/L_V \sim 2.5[M/L_V]_\odot$, then each dSph is embedded in a dark

¹⁰Chanamé et al. (2008) have formulated a Schwarzschild method that operates on discrete velocity measurements, avoiding the binning procedure altogether. Efforts to apply this method to dSph data are underway.



■ Fig. 20-19

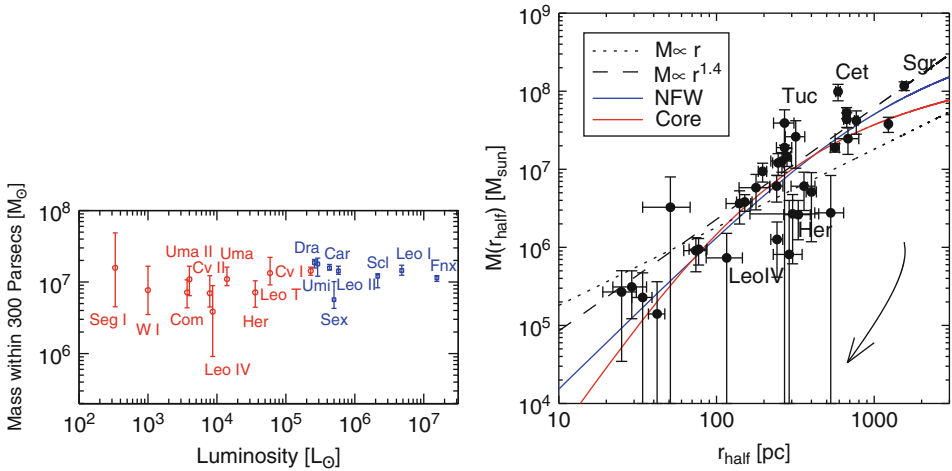
Empirical constraints on the slopes of mass profiles in Fornax and Sculptor based on estimates of $M(r_h)$ for each of two chemodynamically independent stellar subpopulations in each galaxy (Sect. 4.2.3, Walker and Peñarrubia 2011, reproduced by permission of the American Astronomical Society). Points in the left two panels indicate constraints on the two half-light radii and masses enclosed therein, with lines indicating the maximum slopes allowed for cored ($\gamma = 0$ in the notation of 20.12); $\Gamma \equiv \Delta \log M / \Delta \log r \leq 3$) and cusped ($\gamma = 1$, $\Gamma \leq 2$) overplotted. The right-hand panel indicates posterior probability distributions for the slope in each galaxy. The data rule out the cuspy ($\gamma \geq 1$, $\Gamma \leq 2$) profiles that characterize the cold dark matter halos produced in cosmological simulations (e.g., Navarro et al. 1997) with significance $\gtrsim 96\%$ and $\gtrsim 99\%$ in Fornax and Sculptor, respectively

matter halo of mass $\sim 10^7 M_{\odot}$. Allowing for extended dark matter halos, Mateo et al. (1993) and Mateo (1998) interpret this value as the minimum mass that is associated empirically with dark matter.

Kinematic studies of the newfound ultrafaint satellites alter this picture slightly. The ultrafaint dSphs extend the luminosity floor from $M_V \sim -9$ (e.g., Draco, Ursa Minor) to $M_V \sim -2$ (e.g., Segue 1), nearly three orders of magnitude in luminosity. Applying mass-follows-light models to their velocity data for the least luminous dSphs, Martin et al. (2007) and Simon and Geha (2007) estimate dynamical masses as small as $\sim 10^5 M_{\odot}$. While these masses imply extreme mass-to-light ratios $M/L_V \gtrsim 100 [M/L_V]_{\odot}$ that extend the monotonic increase in M/L_V with decreasing luminosity, they suggest that this relationship becomes flatter toward the smallest luminosities (right panel of Fig. 20-17).

Strigari et al. (2008a) extend to ultrafaint satellites the notion of a common dSph mass by considering the mass enclosed within a fixed radius of 300 pc. Using a Jeans analysis similar to that described in Sect. 4.1.3, Strigari et al. (2008a) estimate $M_{300} \equiv M(r \leq 300 \text{ pc}) \sim 10^7 M_{\odot}$ for dSphs spanning five orders of magnitude in luminosity (Fig. 20-20, left panel).¹¹ Walker et al. (2009b) take a different approach using the model-independent estimates of $M(R_h)$ provided by (20.15) to evaluate the hypothesis that all dSphs are embedded in identical dark matter halos characterized by a “universal” mass profile (Fig. 20-20, right panel). While the

¹¹Implicit in the estimation of M_{300} is the assumption that the smallest dSphs, which have half-light radii $R_h \sim 30 \text{ pc}$, have dark matter halos that extend beyond their most distant dynamical tracers.



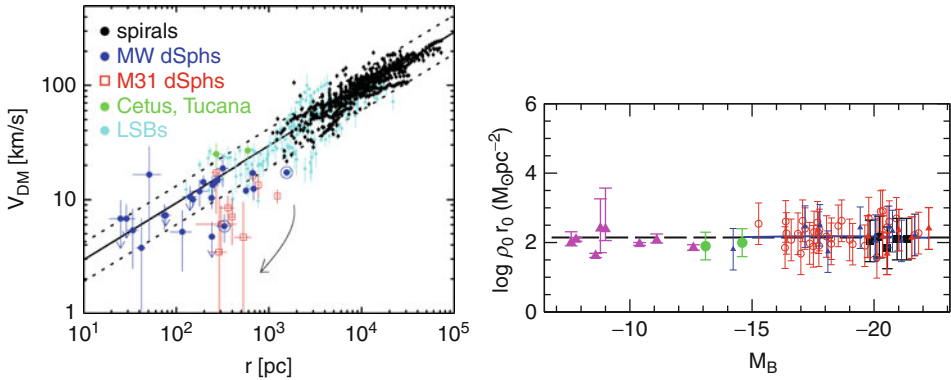
■ Fig. 20-20

A common mass? *Left*: mass enclosed within a radius of $r \leq 300$ pc estimated from a Jeans analysis (See ▶ Sect. 4.1.3) versus luminosity (Strigari et al. 2008a, reproduced with permission). *Right*: mass enclosed within the projected half-light radius, estimated from (▶ 20.15) versus half-light radius (Walker et al. 2009b, reproduced by permission of the American Astronomical Society). Arrows indicate the trajectory followed by satellites as they lose up to 99% of their original stellar mass from N-body simulations by Peñarrubia et al. (2008). Overplotted are various dark matter halo models (Walker et al. 2009b)

scatter of $M(R_h)$ values about a single power law $M(r) \propto r^x$ is larger (by a factor of ~ 2) than the scatter expected to arise from observational errors, it is similar to the scatter about a common value of M_{300} .

While it has long been known that dSphs do not fit naturally onto scaling relations defined by larger elliptical galaxies (e.g., Kormendy 1985), the roughly exponential decline of dSph surface brightness profiles (▶ Sect. 2.2.1) hints at a structural relationship to spirals (Faber and Lin 1983; Kormendy 1985; Lin and Faber 1983), perhaps via an evolutionary mechanism such as tidal stirring (▶ Sects. 2.2.1 and ▶ Sect. 3.2, Mayer et al. 2001a, b). Converting estimates of $M(R_h)$ to circular velocities $v^2 = GM(R_h)/R_h$, Walker et al. (2010) find that the Milky Way's dSph satellites land on an extrapolation of the mean rotation curve estimated for dark matter in spiral galaxies (▶ Fig. 20-21, left panel), $\log_{10}[V_{\text{circ}}/(\text{kms}^{-1})] = 1.5 + 0.5 \log_{10}[r/\text{kpc}]$ (McGaugh et al. 2007). Examining a scaling relation previously identified by Kormendy and Freeman (2004), Donato et al. (2009), and Salucci et al. (2011) find similar results: assuming cored dark matter halos of the form $\rho(r) = \rho_0 r_0^3 (r + r_0)^{-1} (r^2 + r_0^2)^{-1}$ (Berkert 1995), fits to dSph velocity dispersion profiles and spiral rotation curves suggest a single, simple relationship characterized by $\log_{10}[\rho_0/(M_\odot \text{pc}^{-3}) r_0/\text{pc}] \sim 2$ independently of luminosity (▶ Fig. 20-21, right panel).

These scaling relationships have provoked a variety of disparate interpretations. For example, Boyarsky et al. (2010a) show analytically that "secondary infall" of dark matter onto isolated halos would cause surface density to vary slowly with halo mass, perhaps explaining the apparent constancy of $\rho_0 r_0$ over the range spanned by spirals. They argue that the apparent



■ Fig. 20-21

Similarity with spiral galaxies (☛ Sect. 5). *Left*: circular velocities ($v_{\text{circ}} = \sqrt{GM(R_h)/R_h}$) for the Milky Way’s dSph satellites lie on the extrapolation of the mean rotation curve due to dark matter in spiral (McGaugh et al. 2007) and low-surface-brightness (Kuzio de Naray et al. 2008) galaxies (Walker et al. 2010, reproduced by permission of the American Astronomical Society). *Right*: dark matter surface densities inferred from fits of Burkert (1995) halos have the same constant value inferred from spiral galaxy rotation curves (Reproduced from *Dwarf Spheroidal Galaxy Kinematics and Spiral Scaling Laws*, by Salucci et al. (2011), by permission of Wiley)

extension to dSphs agrees with results from cosmological N-body simulations (e.g., Springel et al. 2008), which produce satellite halos with approximately the surface density estimated for the real dSphs. On the other hand, McGaugh and Wolf (2010) show that dSphs deviate from another spiral scaling relation—the baryonic Tully-Fisher relation (“BTF”, McGaugh et al. 2000)—systematically according to the amount by which their half-light radii exceed their tidal radii, provided that the latter radius is estimated using Milgrom’s (1983) Modified Newtonian Dynamics (a framework that naturally implies the BTF relation). In any case, scatter among the Milky Way’s dSphs increased when Adén et al. (2009) and Koposov et al. (2011b) measured smaller velocity dispersions for Hercules and Boötes I, respectively, than had previously been estimated (☛ Sect. 2.3). Furthermore, early kinematic results for the dSph satellites of M31 suggest systematically smaller masses (at a given half-light radius) than estimated for the Milky Way dSphs (Collins et al. 2010, 2011; Kalirai et al. 2010). However, more recent, larger data sets suggest that the M31 dSphs are more similar to the MW dSphs than previously suggested (Tollerud et al. 2011). Pending more detailed spectroscopic surveys, particularly of the least luminous and most distant dSphs, the interpretation of scaling relationships that connect galaxies across such wide ranges of luminosity and morphological type should proceed with caution.

6 Implications for Cosmology and Particle Physics

Observations indicate that the Milky Way’s known dSph satellites have masses $10^5 \lesssim M(R_h)/M_{\odot} \lesssim 10^7$ and that dark matter dominates their internal kinematics at all radii, $M/L_V \gtrsim 10[M/L_V]_{\odot}$ (☛ Sect. 4.1). Current observations also provide direct and/or indirect constraints on the internal distributions of dark matter in three dSphs: Fornax, Sculptor, and

Ursa Minor (● Sect. 4.2). For all three, the available evidence indicates central “cores” of constant density on scales of a few hundred pc. Taken at face value, these basic results have implications for broader areas of physics.

6.1 Cosmology

Observations of structure on large scales (e.g., as inferred from redshift surveys and anisotropy of the cosmic microwave background radiation) seem to require a significant contribution to the mass budget from non-baryonic dark matter, $\Omega_{\text{DM}} \sim 0.22$ (e.g., Bennett et al. 2003; Spergel et al. 2003). The “cold dark matter” (CDM) cosmological paradigm is built on the hypothesis that the dark matter consists of fundamental particles that act like a collisionless gas after decoupling from radiation at nonrelativistic speeds shortly after the Big Bang. Small cross sections and low thermal velocities allow CDM structure to form and survive at high densities in small volumes, thereby enabling the growth of structure on small scales in the early universe.

Calculations of the matter power spectrum associated with popular “weakly interacting massive particle” (WIMP) candidates for the dark matter (e.g., neutralinos with mass $m_\chi \gtrsim 10$ GeV) indicate $P(k) \propto k^{-3}$ at small scales until collisional damping and free streaming finally cause an exponential decline on sub-parsec (comoving) scales (Diemand et al. 2005a; Green et al. 2004). The corresponding halo mass function would be approximately $dN(M)/dM \propto M^{-\alpha}$ with $\alpha \sim -1.9$, and a galaxy like the Milky Way would host roughly $\sim 10^{15}$ satellites in the form of individual, self-bound dark matter “subhalos,” “sub-subhalos,” ..., and “microhalos” with masses $\gtrsim 10^{-6} M_\odot$ (Diemand et al. 2005a; Hofmann et al. 2001; Springel et al. 2008).

Requirements derived from the current census of Milky Way satellites seem rather modest in this context. A viable dark matter particle needs only to accommodate the formation and survival of only a few tens (or hundreds when correcting for incompleteness of sky surveys, Koposov et al. 2008; Tollerud et al. 2008) of dark matter halos with masses $M \gtrsim 10^5 M_\odot$ (● Sect. 4) around the Milky Way. These constraints allow for lighter “warmer” particle candidates (e.g., the sterile neutrino, Dodelson and Widrow 1994) whose longer free-streaming lengths would truncate the matter power spectrum at scales more similar to those that characterize the smallest galaxies (Bode et al. 2001; Gilmore et al. 2007; Lovell et al. 2012; Macciò and Fontanot 2010; Polisensky and Ricotti 2011).

Thus dark matter particle candidates and associated cosmologies can be classified according to whether the particles’ free streaming plays a significant role in galaxy formation (e.g., Bøhm et al. 2001). For sufficiently massive and “cold” particles, it does not, and other physical processes must be invoked to explain the suppression and/or truncation of galaxy formation in low-mass halos (e.g., Font et al. 2011; Klypin et al. 1999; Koposov et al. 2009; Kravtsov 2010; Li et al. 2010; Macciò et al. 2010). The negligible thermal velocities invoked for “standard” CDM particles also imply that N-body simulations can track the growth of structure accurately with relatively few particles, making CDM cosmological simulations the simplest, fastest, and most widely practiced kind.

Cosmological simulations demonstrate that if (standard) gravitational interactions between CDM particles dominate the formation and evolution of galactic structure, then galaxies ought to be embedded in dark matter halos that have central cusps characterized by $\lim_{r \rightarrow 0} \rho(r) \propto r^{-\gamma}$, with $\gamma \gtrsim 1$ (e.g., Diemand et al. 2005b; Dubinski and Carlberg 1991; Klypin et al. 2001; Moore et al. 1998; Navarro et al. 1996b, 1997; Springel et al. 2008). Observations indicate that most individual galaxies are not embedded in such halos. Instead, rotation curves of spiral and low-surface-brightness galaxies tend to favor dark matter halos with resolved “cores” ($\gamma \sim 0$) of

constant density (e.g., de Blok and McGaugh 1997; Flores and Primack 1994; Kuzio de Naray et al. 2006, 2008; McGaugh et al. 2001; Moore 1994; Salucci and Burkert 2000; Simon et al. 2005, de Blok 2010, and references therein). These results imply that (standard) gravitational interactions between CDM particles do not always dominate the formation and evolution of galactic structure.

Indeed, galaxies contain baryons prone to interact via forces other than gravity. Many hydrodynamical simulations demonstrate that various poorly understood baryon-physical mechanisms might influence the structure of galactic CDM halos (e.g., Blumenthal et al. 1986; Del Popolo 2010; El-Zant et al. 2001; Gnedin et al. 2004; Governato et al. 2010; Navarro et al. 1996a; Pontzen and Governato 2011; Romano-Díaz et al. 2009; Tonini et al. 2006). Insofar as their baryons are dynamically negligible, dSphs and low-surface-brightness galaxies enable the most direct comparisons to structures formed in CDM-only simulations. In this context, the available evidence against cusped dark matter halos in Fornax, Sculptor, and Ursa Minor (♣ Sect. 4.2) becomes particularly relevant: the viability of CDM now requires that baryon-driven mechanisms can have *reduced* the central dark matter densities in these galaxies to $\rho_0 \gtrsim 5 \times 10^7 M_\odot \text{kpc}^{-3}$ while leaving behind stellar populations with low luminosities $10^5 \lesssim L_V/L_{V,\odot} \lesssim 10^7$ and central surface brightnesses $23 \lesssim \mu_0/(\text{mag}/\text{arcsec}^2) \lesssim 25$.¹²

Recent work identifies several mechanisms that might accomplish this feat on dSph scales by invoking either the dynamical coupling of the dark matter to energetic baryonic outflows (e.g., de Souza et al. 2011; Mashchenko et al. 2006, 2008; Read and Gilmore 2005) or the transfer of energy/angular momentum to dark matter from massive infalling objects (e.g., Cole et al. 2011; Goerdt et al. 2006, 2010; Sánchez-Salcedo et al. 2006). Hydrodynamical simulations by Sawala et al. (2010) and Parry et al. (2011) indicate that the former category of solutions has difficulty reproducing other dSph observables – specifically, star formation histories as well as luminosity functions and metallicity distributions. The latter category is difficult to evaluate observationally, as the evidence can literally be destroyed (e.g., by tidal disruption); furthermore it seems unlikely that such infall mechanisms generate cores of sufficient size.¹³

In any case, the emerging challenge for the standard CDM paradigm is not that empirical; evidence against cusped dark matter halos necessarily rules out the hypothesis that CDM particles constitute the dark matter. The poorly understood complexities of baryon physics – along with the freedom to invoke other processes, e.g., self-scattering of CDM particles (Loeb and Weiner 2011; Spergel and Steinhardt 2000; Vogelsberger et al. 2012) – leave sufficient flexibility for CDM to be rendered consistent with virtually any realistic observation of galactic structure. In fact that is the problem. CDM escapes falsification of perhaps its most famous prediction only by withdrawing the prediction. While this circumstance does not imply that CDM is incorrect, it does mean that CDM currently fails to make accurate predictions regarding the stellar dynamics of galaxies, a primary piece of evidence for dark matter in the first place. In this context, an outcome favorable to standard CDM seems to require the detection of either (1) gravitational interactions involving dark matter halos on subgalactic scales (e.g., via microlensing or perturbations of loosely bound luminous structure) or (2) nongravitational interactions involving cold dark matter particles.

¹²Boylan-Kolchin et al. (2011) identify a similar (perhaps the same) structural problem, noting that the most massive “subhalos” produced in the *Aquarius* CDM simulation (Springel et al. 2008) have central densities larger than those estimated for any of the known dSphs.

¹³For example, Goerdt et al. (2010) conclude that a sinking object of mass M_s induces core formation inside a radius where the enclosed halo mass is $M(r_{\text{core}}) \sim M_s$. In this scenario, the sinking of Fornax’s five surviving globular clusters ($M_s \sim 10^5 M_\odot$) cannot have formed the core inferred from estimates $M(\sim 550 \text{pc}) \sim 5 \times 10^7 M_\odot$ and $M(\sim 900 \text{pc}) \sim 2 \times 10^8 M_\odot$ (Walker and Peñarrubia 2011).

6.2 Particle Physics

It has long been recognized that the small sizes and large mass densities of dSphs place strong constraints on the particle nature of dark matter. For example, Liouville's theorem requires that in the phase-space densities of light, neutral lepton species do not increase after decoupling from radiation in the early Universe. Tremaine and Gunn (1979) point out that this constraint, combined with the necessity that $\Omega_\nu < \Omega_{\text{matter}}$, places a conservative upper limit on the neutrino mass that is summarily violated by lower limits from phase-space densities inferred for galaxy halos. Therefore, neutrinos are not the dark matter in galaxies. The exclusion of neutrinos is most evident on small scales, where small volumes demand heavy particles in order to satisfy phase-space requirements.

For example, using Aaronson's (1983) initial measurement of Draco's velocity dispersion, Lin and Faber (1983) derive a lower limit of $m_\nu \gtrsim 500$ eV. Lake (1989) points out that this constraint is sensitive to the dubious assumption that mass follows light (► Sect. 4.1.1). Gerhard and Spergel (1992) strengthen the argument by turning it around, noting that for more viable neutrino masses of $m_\nu \sim 30$ eV, the core radii of dSph halos would need to be unrealistically large ($\gtrsim 10$ kpc) to accommodate model-independent lower limits of $\rho_0 \gtrsim 0.05 M_\odot \text{pc}^{-3}$ (Pryor and Kormendy 1990) on their central densities. Finally, generalising the phase-space argument of Tremaine and Gunn (1979) to relativistically decoupled warm dark matter candidates, Dalcanton and Hogan (2001) show that of all galaxies, dSphs provide the most stringent limits, $m_\chi \gtrsim 700$ eV and $m_\chi \gtrsim 30$ eV for thermal and degenerate fermions, respectively. Most recently and more specifically, Boyarsky et al. (2009a) use phase-space arguments to conclude that $m_\chi \gtrsim 1.7$ keV if the dark matter consists of sterile neutrinos produced via non-resonant mixing with active neutrinos.

Any positive identification of a dark matter particle will require the detection of its nongravitational interactions. Experiments at the Large Hadron Collider might find evidence for such interactions, as might various experiments designed to detect directly the scattering of dark matter particles in Earth's orbital path. Alternatively, high-energy photons might be released if dark matter self-annihilates (Gunn et al. 1978; Stecker 1978) or decays (Boyarsky et al. 2006; Kusenko 2006; Pal and Wolfenstein 1982), providing an opportunity for indirect detection.

Their large mass-to-light ratios, low astrophysical backgrounds, and close proximities make the Milky Way's dSph satellites popular targets in the search for annihilation and/or decay products (e.g., Evans et al. 2004; Kuhlen 2010; Strigari et al. 2008b). For annihilation, the differential γ -ray flux (units $\text{cm}^{-2} \text{s}^{-1} \text{sr}^{-1} \text{GeV}^{-1}$) received on Earth in solid angle $\Delta\Omega$ is given by

$$\frac{d\Phi_\gamma}{dE_\gamma} = \frac{1}{4\pi} \frac{\langle\sigma v\rangle}{2m_\chi^2} \cdot \frac{dN_\gamma}{dE_\gamma} \times J(\Delta\Omega), \quad (20.17)$$

where m_χ is the particle mass, $\langle\sigma v\rangle$ is the (velocity averaged) cross section, dN_γ/dE_γ is the energy spectrum of products, and

$$J(\Delta\Omega) = \int_{\Delta\Omega} \int \rho^2(l, \Omega) dl d\Omega. \quad (20.18)$$

This “ J -factor” represents the astrophysical contribution to the signal and is specified by the integral of the squared dark matter density, $\rho^2(l, \Omega)$, over line-of-sight l and solid angle Ω . The equation for the flux due to decay events is similar, except that the integral is taken over the dark matter density raised only to the first power. In practice, constraints on J come directly from

constraints on $\rho(r)$ obtained in parametric Jeans analyses of the sort described in [Sect. 4.1.3](#) and demonstrated in [Fig. 20-18](#) (e.g., Charbonnier et al. 2011; Martinez et al. 2009; Strigari et al. 2007b).

At present, dSph surveys conducted with atmospheric Cherenkov telescopes (e.g., Aleksic et al. 2011; Essig et al. 2009; H. E. S. S. Collaboration et al. 2011; Pieri et al. 2009; The VERITAS collaboration: Vivier et al. 2011), x-ray (e.g., Boyarsky et al. 2007, 2010b; Loewenstein and Kusenko 2010, 2012; Loewenstein et al. 2009), and gamma-ray telescopes (e.g., Abdo et al. 2010; Ackermann et al. 2011) yield no unambiguous detections.¹⁴ From [\(20.17\)](#), upper limits on photon flux translate into upper limits on the cross section $\langle\sigma v\rangle$ for a given particle mass. For example, [Fig. 20-22](#) plots 95% upper limits on $\langle\sigma v\rangle$ derived from Fermi-LAT observations of Milky Way dSphs based on 2 years of data from the planned 5-year mission (Geringer-Sameth and Koushiappas 2011; Ackermann et al. 2011). Dotted lines at $\langle\sigma v\rangle \sim 3 \times 10^{-26} \text{ cm}^3 \text{ s}^{-1}$ indicate the “generic” cross section expected for WIMPs with mass $m_\chi \sim 0.1\text{--}1 \text{ TeV}$ (e.g., Jungman et al. 1996). WIMPs having this combination of mass and cross section would have decoupled from radiation with relic abundance $\Omega_{\text{WIMP}} \sim 0.2$, the value cosmology requires of the dark matter (a coincidence known as the “WIMP miracle,” e.g., Feng 2010; Jungman et al. 1996). For particle masses below $m_\chi \lesssim 10 \text{ GeV}$, the combination of kinematic and high-energy data available for dSphs is now beginning to exclude the cross section most readily associated with WIMPs.¹⁵ Over the next decade, searches for dark matter and/or its by-products will intensify with large-scale efforts at existing facilities and with new instrumentation that will provide unprecedented sensitivity (e.g., CTA Consortium 2010).

The status of dark matter will depend critically on the outcomes of direct and indirect dark matter detection experiments that are either ongoing or planned for the near future. Also at stake is the motivation and context for studying dark matter phenomenology in dSphs. An unambiguous, positive detection of dark matter by-products emitted from dSphs would establish the existence of a new particle and would provide a unique means for measuring its mass and cross section; via [\(20.17\)](#), constraints on such parameters would be only as good as dynamical constraints on the dark matter density profile. In case of detections in other objects, e.g., the Galactic center (see Hooper and Linden 2011) or galaxy clusters (see Han et al. 2012),

¹⁴Loewenstein and Kusenko (2010) interpret a *Chandra* detection of monochromatic ($\sim 2.5 \text{ keV}$) emission from the direction of the Willman 1 satellite as a decay signal. However, Boyarsky et al. (2010b) argue that non-detections of this feature in the Galactic halo, M31, and several other dSphs rule out such an interpretation. Indeed, Loewenstein and Kusenko (2012) report no detection of the $\sim 2.5 \text{ keV}$ feature in follow-up XMM-Newton observations of Willman 1; corresponding limits on the mass/mixing angle of sterile neutrinos depend on how reliably the “irregular” stellar kinematics of Willman 1 (Willman et al. 2011, [Section 2.2.4](#)) trace its mass.

¹⁵Charbonnier et al. (2011) use published kinematic data to estimate less stringent limits of $\langle\sigma v\rangle \lesssim 10^{-25} \text{ cm}^3 \text{ s}^{-1}$ (at $m_\chi \sim 10 \text{ GeV}$, cf. [Figure 20-22](#)) for individual dSphs. Possible reasons for this discrepancy include different assumptions about the dark matter halo profile (Geringer-Sameth and Koushiappas 2011 and Ackermann et al. 2011 adopt J values previously estimated under the assumption that dSph dark matter halos follow NFW profiles; Charbonnier et al. 2011 estimate J values by marginalizing over uncertain halo shape parameters), different assumptions about the energy spectrum (Geringer-Sameth and Koushiappas 2011 and Ackermann et al. 2011 explicitly consider annihilation via $b\bar{b}$ and $\tau^+\tau^-$ mechanisms; Charbonnier et al. 2011 consider a conservative spectrum averaged over a variety of plausible annihilation channels) and/or different assumptions about detector sensitivity.

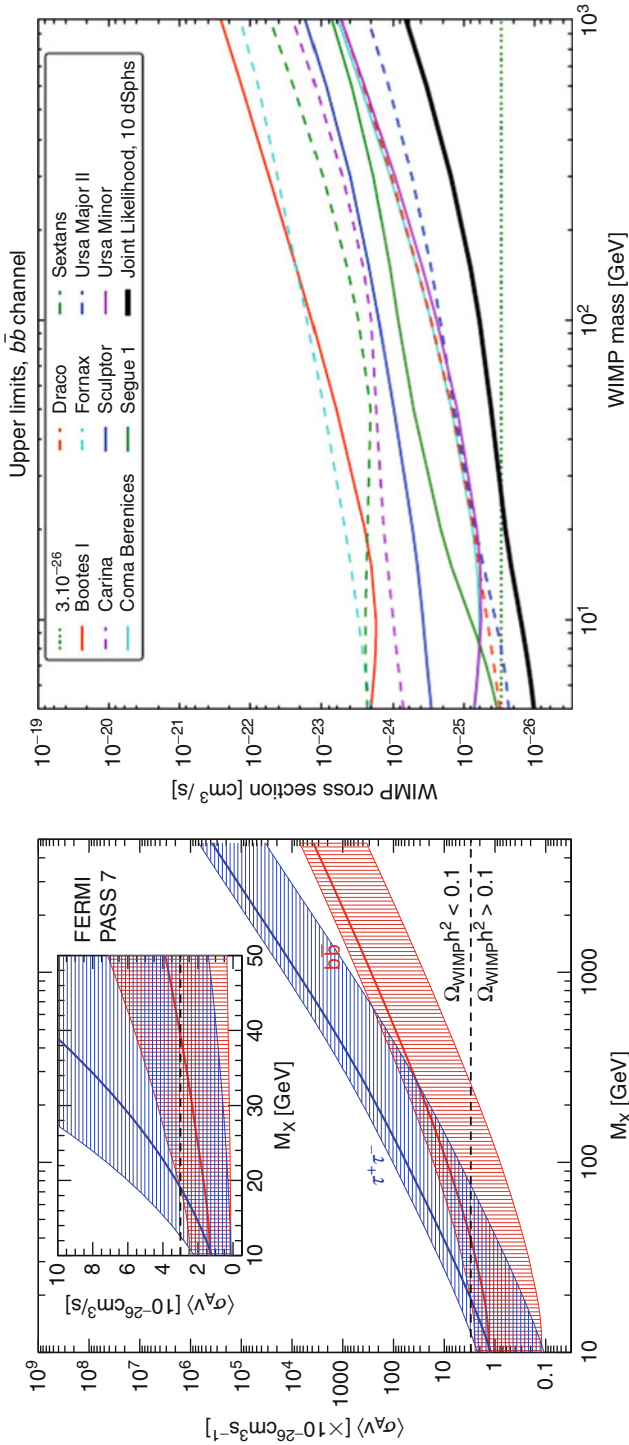


Fig. 20-22 Exclusion of WIMP self-annihilation cross sections, based on Fermi-LAT non-detections (2-year data) of gamma-rays in the Milky Way's dSph satellites (Reprinted with permission from Geringer-Sameth and Koushiappas (left-hand panel, 2011, Phys. Rev. Lett. 107, 241303) and Ackermann et al. (The Fermi Collaboration, right-hand panel, 2011, Phys. Rev. Lett. 107, 241302), Copyright 2011 by the American Physical Society)

high-energy and dynamical constraints from dSphs would provide important consistency checks in the regime of minimal astrophysical background. In the case of direct detection and characterisation of a new particle in the laboratory, dSph phenomenology would help to establish or rule out association with cosmological dark matter. In case of failure to detect nongravitational interactions involving dark matter particles, consideration of the extreme phenomenology exhibited by dSphs would help to inform alternative explanations for dark matter.

7 Some Considerations for Future Work

Topics and/or questions for future investigations include the following:

- The faintest known satellites are often disproportionately interesting. What sorts of bound stellar/gaseous objects will the next generation of sky surveys detect?
- All dSphs clearly deviate from the simple models invoked to characterize their observed properties. In order to understand dSphs as galaxies, one must consider not only the structural and kinematic data emphasized above but also chemical abundances, star formation histories (i.e., *stellar* masses), internal substructure, and external environment.
- The outer stellar structures of dSphs carry information about the gravitational competition between the Milky Way and its satellites. Combined with Gaia and Gaia-ESO structural/kinematic data for the Milky Way, deep, wide-field spectroscopic surveys that reach large galactocentric distances in dSphs will reveal details of the transition from satellite to host potential.
- Methods for data analysis, mass modeling and mass measurements continue to develop alongside efforts to exploit more of the information contained in available dSph data sets. Promising strategies include the formulation of statistical likelihood functions that depend on discrete measurements (e.g., of individual stellar positions and velocities) rather than on binned profiles and exploitation of any substructure present in the data.
- Can observations and/or experiments provide clear evidence for nongravitational interactions involving a dark matter particle? Can observations provide evidence for gravitational interactions involving dark matter halos on subgalactic scales?
- Are there cosmological and/or particle physics models for dark matter from which accurate predictions regarding the stellar dynamics of the most dark-matter dominated galaxies can be extracted?

The author is grateful to Gerry Gilmore for the invitation to write this review and for insightful comments. The author thanks Giuseppina Battaglia, Alexey Boyarsky, Alexander Kusenko, Oleg Ruchayskiy and Ed Olszewski for suggestions that improved the quality of this work. The author is currently supported by NASA through Hubble Fellowship grant HST-HF-51283.01-A, awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555.

Cross-References

- [History of Dark Matter in Galaxies](#)

References

- Aaronson, M. 1983, *ApJ*, 266, L11
- Aaronson, M., & Olszewski, E. 1987a, in *IAU Symposium, Dark matter in the Universe*, Vol. 117, ed. J. Kormendy, & G. R. Knapp (Dordrecht: Reidel), 153–158
- Aaronson, M., & Olszewski, E. W. 1987b, *AJ*, 94, 657
- Abdo, A. A., Ackermann, M., Ajello, M., Atwood, W. B., Baldini, L., Ballet, J., Barbiellini, G., Bastieri, D., Bechtol, K., Bellazzini, R., Berenji, B., Bloom, E. D., Bonamente, E., Borgland, A. W., Bregeon, J., Brez, A., Brigida, M., Bruel, P., Burnett, T. H., Buson, S., Caliandro, G. A., Cameron, R. A., Caraveo, P. A., Casandjian, J. M., Cecchi, C., Chekhtman, A., Cheung, C. C., Chiang, J., Ciprini, S., Claus, R., Cohen-Tanugi, J., Conrad, J., de Angelis, A., de Palma, F., Digel, S. W., Silva, E. d. C. e., Drell, P. S., Drlica-Wagner, A., Dubois, R., Dumora, D., Farnier, C., Favuzzi, C., Fegan, S. J., Focke, W. B., Fortin, P., Frailis, M., Fukazawa, Y., Fusco, P., Gargano, F., Gehrels, N., Germani, S., Giebels, B., Giglietto, N., Giordano, F., Glanzman, T., Godfrey, G., Grenier, I. A., Grove, J. E., Guillemot, L., Guiriec, S., Gustafsson, M., Harding, A. K., Hays, E., Horan, D., Hughes, R. E., Jackson, M. S., Jeltema, T. E., Jóhannesson, G., Johnson, A. S., Johnson, R. P., Johnson, W. N., Kamae, T., Katagiri, H., Kataoka, J., Kerr, M., Knödlseeder, J., Kuss, M., Lande, J., Latronico, L., Lemoine-Goumard, M., Longo, F., Loparco, F., Lott, B., Lovellette, M. N., Lubrano, P., Madejski, G. M., Makeev, A., Mazziotta, M. N., McEnery, J. E., Meurer, C., Michelson, P. F., Mitthumsiri, W., Mizuno, T., Moiseev, A. A., Monte, C., Monzani, M. E., Moretti, E., Morselli, A., Moskaleiko, I. V., Murgia, S., Nolan, P. L., Norris, J. P., Nuss, E., Ohsugi, T., Omodei, N., Orlando, E., Ormes, J. F., Paneque, D., Panetta, J. H., Parent, D., Pelassa, V., Pepe, M., Pesce-Rollins, M., Piron, F., Porter, T. A., Profumo, S., Rainò, S., Rando, R., Razzano, M., Reimer, A., Reimer, O., Reposeur, T., Ritz, S., Rodriguez, A. Y., Roth, M., Sadrozinski, H. F.-W., Sander, A., Saz Parkinson, P. M., Scargle, J. D., Schalk, T. L., Sellerholm, A., Sgrò, C., Siskind, E. J., Smith, D. A., Smith, P. D., Spandre, G., Spinelli, P., Strickman, M. S., Suson, D. J., Takahashi, H., Takahashi, T., Tanaka, T., Thayer, J. B., Thayer, J. G., Thompson, D. J., Tibaldo, L., Torres, D. F., Tramacere, A., Uchiyama, Y., Usher, T. L., Vasileiou, V., Vilchez, N., Vitale, V., Waite, A. P., Wang, P., Winer, B. L., Wood, K. S., Ylienen, T., Ziegler, M., Bullock, J. S., Kaplinghat, M., Martinez, G. D., & Fermi-LAT Collaboration. 2010, *ApJ*, 712, 147
- Ackermann, M., Ajello, M., Albert, A., Atwood, W. B., Baldini, L., Ballet, J., Barbiellini, G., Bastieri, D., Bechtol, K., Bellazzini, R., Berenji, B., Blandford, R. D., Bloom, E. D., Bonamente, E., Borgland, A. W., Bregeon, J., Brigida, M., Bruel, P., Buehler, R., Burnett, T. H., Buson, S., Caliandro, G. A., Cameron, R. A., Cañadas, B., Caraveo, P. A., Casandjian, J. M., Cecchi, C., Charles, E., Chekhtman, A., Chiang, J., Ciprini, S., Claus, R., Cohen-Tanugi, J., Conrad, J., Cutini, S., de Angelis, A., de Palma, F., Dermer, C. D., Digel, S. W., Do Couto E Silva, E., Drell, P. S., Drlica-Wagner, A., Falletti, L., Favuzzi, C., Fegan, S. J., Ferrara, E. C., Fukazawa, Y., Funk, S., Fusco, P., Gargano, F., Gasparrini, D., Gehrels, N., Germani, S., Giglietto, N., Giordano, F., Giroletti, M., Glanzman, T., Godfrey, G., Grenier, I. A., Guiriec, S., Gustafsson, M., Hadasch, D., Hayashida, M., Hays, E., Hughes, R. E., Jeltema, T. E., Jóhannesson, G., Johnson, R. P., Johnson, A. S., Kamae, T., Katagiri, H., Kataoka, J., Knödlseeder, J., Kuss, M., Lande, J., Latronico, L., Lionetto, A. M., Llena Garde, M., Longo, F., Loparco, F., Lott, B., Lovellette, M. N., Lubrano, P., Madejski, G. M., Mazziotta, M. N., McEnery, J. E., Mehault, J., Michelson, P. F., Mitthumsiri, W., Mizuno, T., Monte, C., Monzani, M. E., Morselli, A., Moskaleiko, I. V., Murgia, S., Naumann-Godo, M., Norris, J. P., Nuss, E., Ohsugi, T., Okumura, A., Omodei, N., Orlando, E., Ormes, J. F., Ozaki, M., Paneque, D., Parent, D., Pesce-Rollins, M., Pierbattista, M., Piron, F., Pivato, G., Porter, T. A., Profumo, S., Rainò, S., Razzano, M., Reimer, A., Reimer, O., Ritz, S., Roth, M., Sadrozinski, H. F.-W., Sbarra, C., Scargle, J. D., Schalk, T. L., Sgrò, C., Siskind, E. J., Spandre, G., Spinelli, P., Strigari, L., Suson, D. J., Tajima, H., Takahashi, H., Tanaka, T., Thayer, J. G., Thayer, J. B., Thompson, D. J., Tibaldo, L., Tinivella, M., Torres, D. F., Troja, E., Uchiyama, Y., Vandenbroucke, J., Vasileiou, V., Vianello, G., Vitale, V., Waite, A. P., Wang, P., Winer, B. L., Wood, K. S., Wood, M., Yang, Z., Zimmer, S., Kaplinghat, M., & Martinez, G. D. 2011, *Phys. Rev. Lett.*, 107, 241302
- Adén, D., Wilkinson, M. I., Read, J. I., Feltzing, S., Koch, A., Gilmore, G. F., Grebel, E. K., & Lundström, I. 2009, *ApJ*, 706, L150
- Adén, D., Eriksson, K., Feltzing, S., Grebel, E. K., Koch, A., & Wilkinson, M. I. 2011, *A&A*, 525, A153

- Aleksić, J., Alvarez, E. A., Antonelli, L. A., Antoranz, P., Asensio, M., Backes, M., Barrio, J. A., Bastieri, D., Becerra González, J., Bednarek, W., Berdyugin, A., Berger, K., Bernardini, E., Biland, A., Blanch, O., Bock, R. K., Boller, A., Bonnoli, G., Borla Tridon, D., Braun, I., Bretz, T., Cañellas, A., Carmona, E., Carosi, A., Colin, P., Colombo, E., Contreras, J. L., Cortina, J., Cossio, L., Covino, S., Dazzi, F., De Angelis, A., De Cea del Pozo, E., De Lotto, B., Delgado Mendez, C., Diago Ortega, A., Doert, M., Domínguez, A., Dominis Prester, D., Dorner, D., Doro, M., Elsaesser, D., Ferenc, D., Fonseca, M. V., Font, L., Fruck, C., García López, R. J., Garczarczyk, M., Garrido, D., Giavitto, G., Godinović, N., Hadasch, D., Häfner, D., Herrero, A., Hildebrand, D., Höhne-Mönch, D., Hose, J., Hrupec, D., Huber, B., Jogler, T., Klepser, S., Krähenbühl, T., Krause, J., La Barbera, A., Lelas, D., Leonardo, E., Lindfors, E., Lombardi, S., López, M., Lorenz, E., Makariev, M., Maneva, G., Mankuzhiyil, N., Mannheim, K., Maraschi, L., Mariotti, M., Martínez, M., Mazin, D., Meucci, M., Miranda, J. M., Mirzoyan, R., Miyamoto, H., Moldón, J., Moralejo, A., Munar-Androver, P., Nieto, D., Nilsson, K., Orito, R., Oya, I., Paiano, S., Paneque, D., Paoletti, R., Pardo, S., Paredes, J. M., Partini, S., Pasanen, M., Paus, F., Perez-Torres, M. A., Persic, M., Peruzzo, L., Pilia, M., Pochon, J., Prada, F., Prada Moroni, P. G., Prandini, E., Puljak, I., Reichardt, I., Reinthal, R., Rhode, W., Ribó, M., Rico, J., Rügamer, S., Saggion, A., Saito, K., Saito, T. Y., Salvati, M., Satalecka, K., Scalzotto, V., Scapin, V., Schultz, C., Schweizer, T., Shayduk, M., Shore, S. N., Silanpää, A., Sitarek, J., Sobczynska, D., Spanier, F., Spiro, S., Stamerra, A., Steinke, B., Storz, J., Strah, N., Surić, T., Takalo, L., Takami, H., Tavecchio, F., Temnikov, P., Terzić, T., Tescaro, D., Teshima, M., Thom, M., Tibolla, O., Torres, D. F., Treves, A., Vankov, H., Vogler, P., Wagner, R. M., Weitzel, Q., Zabalza, V., Zandanel, F., Zanin, R., Fornasa, M., Essig, R., Sehgal, N., & Strigari, L. E. 2011, *JCAP*, 6, 35
- Amorisco, N. C., & Evans, N. W. 2012, *MNRAS*, 419, 184
- Armandroff, T. E., & Da Costa, G. S. 1986, *AJ*, 92, 777
- Armandroff, T. E., Olszewski, E. W., & Pryor, C. 1995, *AJ*, 110, 2131
- Battaglia et al. 2006, *A&A*, 459, 423
- Battaglia, G., Helmi, A., Tolstoy, E., Irwin, M., Hill, V., & Jablonka, P. 2008, *ApJ*, 681, L13
- Battaglia, G., Tolstoy, E., Helmi, A., Irwin, M., Parisi, P., Hill, V., & Jablonka, P. 2011, *MNRAS*, 411, 1013
- Bekenstein, J. D. 2004, *Phys. Rev. D*, 70, 083509
- Bell, E. F., Slater, C. T., & Martin, N. F. 2011, *ApJ*, 742, L15
- Belokurov, V., Zucker, D. B., Evans, N. W., Wilkinson, M. I., Irwin, M. J., Hodgkin, S., Bramich, D. M., Irwin, J. M., Gilmore, G., Willman, B., Vidrih, S., Newberg, H. J., Wyse, R. F. G., Fellhauer, M., Hewett, P. C., Cole, N., Bell, E. F., Beers, T. C., Rockosi, C. M., Yanny, B., Grebel, E. K., Schneider, D. P., Lupton, R., Barentine, J. C., Brewington, H., Brinkmann, J., Harvanek, M., Kleinman, S. J., Krzesinski, J., Long, D., Nitta, A., Smith, J. A., & Snedden, S. A. 2006a, *ApJ*, 647, L111
- Belokurov et al. 2006b, *ApJ*, 642, L137
- Belokurov et al. 2007, *ApJ*, 654, 897
- Belokurov et al. 2008, *ApJ*, 686, L83
- Belokurov, V., Walker, M. G., Evans, N. W., Gilmore, G., Irwin, M. J., Mateo, M., Mayer, L., Olszewski, E., Bechtold, J., & Pickering, T. 2009, *MNRAS*, 397, 1748
- Belokurov, V., Walker, M. G., Evans, N. W., Gilmore, G., Irwin, M. J., Just, D., Kuposov, S., Mateo, M., Olszewski, E., Watkins, L., & Wyrzykowski, L. 2010, *ApJ*, 712, L103
- Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Spergel, D. N., Tucker, G. S., Wollack, E., Wright, E. L., Barnes, C., Greason, M. R., Hill, R. S., Komatsu, E., Nolte, M. R., Odegard, N., Peiris, H. V., Verde, L., & Weiland, J. L. 2003, *ApJS*, 148, 1
- Bergmann, P. G. 1968, *Int. J. Theor. Phys.*, 1, 25
- Binney, J., & Tremaine, S. 2008, *Galactic Dynamics: Second Edition* (Princeton: Princeton University Press)
- Blumenthal, G. R., Faber, S. M., Flores, R., & Primack, J. R. 1986, *ApJ*, 301, 27
- Bode, P., Ostriker, J. P., & Turok, N. 2001, *ApJ*, 556, 93
- Boyarsky, A., Ruchayskiy, O., and Iakubovskiy, D. 2009a, *JCAP*, 3, 5
- Boehm, C., Fayet, P., & Schaeffer, R. 2001, *Phys. Lett. B*, 518, 8
- Boyarsky, A., Neronov, A., Ruchayskiy, O., Shaposhnikov, M., & Tkachev, I. 2006, *Phys. Rev. Lett.*, 97, 261302
- Boyarsky, A., Neronov, A., Ruchayskiy, O., & Tkachev, I. 2010a, *Phys. Rev. Lett.*, 104, 191301
- Boyarsky, A., Ruchayskiy, O., Iakubovskiy, D., Walker, M. G., Riemer-Sørensen, S., & Hansen, S. H. 2010b, *MNRAS*, 407, 1188
- Boyarsky, A., Nevalainen, J., & Ruchayskiy, O. 2007, *A&A*, 471, 51
- Boylan-Kolchin, M., Bullock, J. S., & Kaplinghat, M. 2011, *MNRAS*, 415, L40

- Burkert, A. 1995, *ApJ*, 447, L25
- Cannon, R. D., Hawarden, T. G., & Tritton, S. B. 1977, *MNRAS*, 180, 81P
- Carlin, J. L., Grillmair, C. J., Muñoz, R. R., Nidever, D. L., & Majewski, S. R. 2009, *ApJ*, 702, L9
- Chanamé, J., Kleyna, J., & van der Marel, R. 2008, *ApJ*, 682, 841
- Charbonnier, A., Combet, C., Daniel, M., Funk, S., Hinton, J. A., Maurin, D., Power, C., Read, J. I., Sarkar, S., Walker, M. G., & Wilkinson, M. I. 2011, *MNRAS*, 418, 1526
- Cole, D. R., Dehnen, W., & Wilkinson, M. I. 2011, *MNRAS*, 416, 1118
- Coleman, M., Da Costa, G. S., Bland-Hawthorn, J., Martínez-Delgado, D., Freeman, K. C., & Malin, D. 2004, *AJ*, 127, 832
- Coleman, M. G., Da Costa, G. S., & Bland-Hawthorn, J. 2005a, *AJ*, 130, 1065
- Coleman, M. G., Da Costa, G. S., Bland-Hawthorn, J., & Freeman, K. C. 2005b, *AJ*, 129, 1443
- Coleman et al. 2007, *ApJ*, 668, L43
- Collins, M. L. M., Chapman, S. C., Irwin, M. J., Martin, N. F., Ibata, R. A., Zucker, D. B., Blain, A., Ferguson, A. M. N., Lewis, G. F., McConnachie, A. W., & Peñarrubia, J. 2010, *MNRAS*, 407, 2411
- Collins, M. L. M., Chapman, S. C., Rich, R. M., Irwin, M. J., Peñarrubia, J., Ibata, R. A., Arimoto, N., Brooks, A. M., Ferguson, A. M. N., Lewis, G. F., McConnachie, A. W., & Venn, K. 2011, *MNRAS*, 417, 1170
- CTA Consortium. 2010, *ArXiv:1008.3703*
- Da Costa, G. S., Hatzidimitriou, D., Irwin, M. J., & McMahon, R. G. 1991, *MNRAS*, 249, 473
- Dalcanton, J. J., & Hogan, C. J. 2001, *ApJ*, 561, 35
- de Blok, W. J. G. 2010, *Adv. Astron.*, 2010, 1–4
- de Blok, W. J. G., & McGaugh, S. S. 1997, *MNRAS*, 290, 533
- de Jong, J. T. A., Martin, N. F., Rix, H.-W., Smith, K. W., Jin, S., & Macciò, A. V. 2010, *ApJ*, 710, 1664
- de Souza, R. S., Rodrigues, L. F. S., Ishida, E. E. O., & Opher, R. 2011, *MNRAS*, 415, 2969
- de Vaucouleurs, G. 1948, *Annales d'Astrophysique*, 11, 247
- Del Popolo, A. 2010, *MNRAS*, 408, 1808
- Diemand, J., Moore, B., & Stadel, J. 2005a, *Nature*, 433, 389
- Diemand, J., Zemp, M., Moore, B., Stadel, J., & Carollo, C. M. 2005b, *MNRAS*, 364, 665
- Dodelson, S., & Widrow, L. M. 1994, *Phys. Rev. Lett.*, 72, 17
- Donato, F., Gentile, G., Salucci, P., Frigerio Martins, C., Wilkinson, M. I., Gilmore, G., Grebel, E. K., Koch, A., & Wyse, R. 2009, *MNRAS*, 397, 1169
- Dubinski, J., & Carlberg, R. G. 1991, *ApJ*, 378, 496
- Duquenoey, A., & Mayor, M. 1991, *A&A*, 248, 485
- El-Zant, A., Shlosman, I., & Hoffman, Y. 2001, *ApJ*, 560, 636
- Essig, R., Sehgal, N., & Strigari, L. E. 2009, *Phys. Rev. D*, 80, 023506
- Evans, N. W., Ferrer, F., & Sarkar, S. 2004, *Phys. Rev. D*, 69, 123501
- Evans, N. W., An, J., & Walker, M. G. 2009, *MNRAS*, 393, L50
- Faber, S. M., & Lin, D. N. C. 1983, *ApJ*, 266, L17
- Fabrizio, M., Nonino, M., Bono, G., Ferraro, L., François, P., Iannicola, G., Monelli, M., Thévenin, F., Stetson, P. B., Walker, A. R., Buonanno, R., Caputo, F., Corsi, C. E., Dall'Orta, M., Gilmozzi, R., James, C. R., Merle, T., Pulone, L., & Romaniello, M. 2011, *PASP*, 123, 384
- Feng, J. L. 2010, *ARA&A*, 48, 495
- Fleck, J.-J., & Kuhn, J. R. 2003, *ApJ*, 592, 147
- Flores, R. A., & Primack, J. R. 1994, *ApJ*, 427, L1
- Font, A. S., Benson, A. J., Bower, R. G., Frenk, C. S., Cooper, A., De Lucia, G., Helly, J. C., Helmi, A., Li, Y.-S., McCarthy, I. G., Navarro, J. F., Springel, V., Starkeburg, E., Wang, J., & White, S. D. M. 2011, *MNRAS*, 417, 1260
- Geha, M., Willman, B., Simon, J. D., Strigari, L. E., Kirby, E. N., Law, D. R., & Strader, J. 2009, *ApJ*, 692, 1464
- Gerhard, O. E., & Spergel, D. N. 1992, *ApJ*, 389, L9
- Geringer-Sameth, A., & Koushiappas, S. M. 2011, *Phys. Rev. Lett.*, 107, 241303
- Gilmore, G., Wilkinson, M. I., Wyse, R. F. G., Kleyna, J. T., Koch, A., Evans, N. W., & Grebel, E. K. 2007, *ApJ*, 663, 948
- Gnedin, O. Y., Hernquist, L., & Ostriker, J. P. 1999, *ApJ*, 514, 109
- Gnedin, O. Y., Kravtsov, A. V., Klypin, A. A., & Nagai, D. 2004, *ApJ*, 616, 16
- Goerdt, T., Moore, B., Read, J. I., Stadel, J., & Zemp, M. 2006, *MNRAS*, 368, 1073
- Goerdt, T., Moore, B., Read, J. I., & Stadel, J. 2010, *ApJ*, 725, 1707
- Governato, F., Brook, C., Mayer, L., Brooks, A., Rhee, G., Wadsley, J., Jonsson, P., Willman, B., Stinson, G., Quinn, T., & Madau, P. 2010, *Nature*, 463, 203
- Green, A. M., Hofmann, S., & Schwarzs, D. J. 2004, *MNRAS*, 353, L23
- Grillmair, C. J. 2009, *ApJ*, 693, 1118
- Gunn, J. E., & Griffin, R. F. 1979, *AJ*, 84, 752
- Gunn, J. E., Lee, B. W., Lerche, I., Schramm, D. N., & Steigman, G. 1978, *ApJ*, 223, 1015
- H. E. S. S. Collaboration, Abramowski, A., Acero, F., Aharonian, F., Akhperjanian, A. G., Anton, G., Barnacka, A., Barres de Almeida, U., Bazer-Bachi, A. R., Becherini, Y., Becker, J., Behera, B., Bernlöhr, K., Bochow, A., Boisson, C., Bolmont, J., Bordas, P., Borrel, V., Brucker, J., Brun, F., Brun, P., Bulik, T., Büsching, I., Carrigan,

- S., Casanova, S., Cerruti, M., Chadwick, P. M., Charbonnier, A., Chaves, R. C. G., Cheesebrough, A., Chounet, L.-M., Clapson, A. C., Coignet, G., Conrad, J., Dalton, M., Daniel, M. K., Davids, I. D., Degrange, B., Deil, C., Dickinson, H. J., Djannati-Ataï, A., Domainko, W., Drury, L. O. C., Dubois, F., Dubus, G., Dyks, J., Dyrda, M., Egberts, K., Eger, P., Espigat, P., Fallon, L., Farnier, C., Fegan, S., Feinstein, F., Fernandes, M. V., Fiasson, A., Fontaine, G., Förster, A., Fülling, M., Gallant, Y. A., Gast, H., Gérard, L., Gerbig, D., Giebels, B., Glicenstein, J. F., Glück, B., Goret, P., Göring, D., Hague, J. D., Hampf, D., Hauser, M., Heinz, S., Heinzelmann, G., Henri, G., Hermann, G., Hinton, J. A., Hoffmann, A., Hofmann, W., Hofverberg, P., Horns, D., Jacholkowska, A., de Jager, O. C., Jahn, C., Jamrozy, M., Jung, I., Kastendieck, M. A., Katarzyński, K., Katz, U., Kaufmann, S., Keogh, D., Kerschhaggl, M., Khangulyan, D., Khélifi, B., Klochkov, D., Kluźniak, W., Kneiske, T., Komin, N., Kosack, K., Kossakowski, R., Laffon, H., Lamanna, G., Lennarz, D., Lohse, T., Lopatin, A., Lu, C.-C., Marandon, V., Marcowith, A., Masbou, J., Maurin, D., Maxted, N., McComb, T. J. L., Medina, M. C., Méhault, J., Moderski, R., Moulin, E., Naumann, C. L., Naumann-Godo, M., de Naurois, M., Nedbal, D., Nekrasov, D., Nguyen, N., Nicholas, B., Niemiec, J., Nolan, S. J., Ohm, S., Olive, J.-F., de Oña Wilhelmi, E., Opitz, B., Ostrowski, M., Panter, M., Paz Arribas, M., Pedalletti, G., Pelletier, G., Petrucci, P.-O., Pita, S., Pühlhofer, G., Punch, M., Quirrenbach, A., Raue, M., Rayner, S. M., Reimer, A., Reimer, O., Renaud, M., de Los Reyes, R., Rieger, F., Ripken, J., Rob, L., Rosier-Lees, S., Rowell, G., Rudak, B., Rulten, C. B., Ruppel, J., Ryde, F., Sahakian, V., Santangelo, A., Schlickeiser, R., Schöck, F. M., Schönwald, A., Schwanke, U., Schwarzbarg, S., Schwemmer, S., Shalchi, A., Sikora, M., Skilton, J. L., Sol, H., Spengler, G., Stawarz, Ł., Steenkamp, R., Stegmann, C., Stinzling, F., Sushch, I., Szostek, A., Tavernet, J.-P., Terrier, R., Tibolla, O., Tluczykont, M., Valerius, K., van Eldik, C., Vasileiadis, G., Venter, C., Vialle, J. P., Viana, A., Vincent, P., Vivier, M., Völk, H. J., Volpe, F., Vorobiov, S., Vorster, M., Wagner, S. J., Ward, M., Wiercholska, A., Zajczyk, A., Zdziarski, A. A., Zech, A., & Zechlin, H.-S. 2011, *Astropart. Phys.*, 34, 608
- Han, J., Frenk, C. S., Eke, V. R., Gao, L., & White, S. D. M. 2012, *ArXiv:1201.1003*
- Hargreaves, J. C., Gilmore, G., Irwin, M. J., & Carter, D. 1994a, *MNRAS*, 269, 957
- Hargreaves, J. C., Gilmore, G., Irwin, M. J., & Carter, D. 1994b, *MNRAS*, 271, 693
- Hargreaves, J. C., Gilmore, G., & Annan, J. D. 1996a, *MNRAS*, 279, 108
- Hargreaves, J. C., Gilmore, G., Irwin, M. J., & Carter, D. 1996b, *MNRAS*, 282, 305
- Harrington, R. G., & Wilson, A. G. 1950, *PASP*, 62, 118
- Harris, W. E. 1996, *AJ*, 112, 1487
- Heisler, J., Tremaine, S., & Bahcall, J. N. 1985, *ApJ*, 298, 8
- Hernandez, X., & Gilmore, G. 1998, *MNRAS*, 297, 517
- Hernquist, L. 1990, *ApJ*, 356, 359
- Hodge, P. W. 1961a, *AJ*, 66, 384
- Hodge, P. W. 1961b, *AJ*, 66, 249
- Hodge, P. W. 1962, *AJ*, 67, 125
- Hodge, P. W. 1963, *AJ*, 68, 470
- Hodge, P. W. 1964a, *AJ*, 69, 853
- Hodge, W. P. 1964b, *AJ*, 69, 438
- Hodge, P. W. 1966, *ApJ*, 144, 869
- Hodge, P. W., & Michie, R. W. 1969, *AJ*, 74, 587
- Hofmann, S., Schwarz, D. J., & Stöcker, H. 2001, *Phys. Rev. D*, 64, 083507
- Hooper, D., & Linden, T. 2011, *Phys. Rev. D*, 83, 083517
- Hubble, E. P. 1930, *ApJ*, 71, 231
- Huxor, A. P., Tanvir, N. R., Irwin, M. J., Ibata, R., Collett, J. L., Ferguson, A. M. N., Bridges, T., & Lewis, G. F. 2005, *MNRAS*, 360, 1007
- Ibata, R. A., Gilmore, G., & Irwin, M. J. 1994, *Nature*, 370, 194
- Ibata, R. A., Wyse, R. F. G., Gilmore, G., Irwin, M. J., & Suntzeff, N. B. 1997, *AJ*, 113, 634
- Ibata et al. 2007, *ApJ*, 671, 1591
- Illingworth, G. 1976, *ApJ*, 204, 73
- Irwin, M., & Hatzidimitriou, D. 1995, *MNRAS*, 277, 1354
- Irwin, M. J., Bunclark, P. S., Bridgeland, M. T., & McMahon, R. G. 1990, *MNRAS*, 244, 16P
- Irwin et al. 2007, *ApJ*, 656, L13
- Irwin, M. J., Ferguson, A. M. N., Huxor, A. P., Tanvir, N. R., Ibata, R. A., & Lewis, G. F. 2008, *ApJ*, 676, L17
- Jardel, J., & Gebhardt, K. 2011, *ArXiv:1112.0319*
- Johnston, K. V., Law, D. R., & Majewski, S. R. 2005, *ApJ*, 619, 800
- Jungman, G., Kamionkowski, M., & Griest, K. 1996, *Phys. Rep.*, 267, 195
- Kalirai, J. S., Beaton, R. L., Geha, M. C., Gilbert, K. M., Guhathakurta, P., Kirby, E. N., Majewski, S. R., Osthimer, J. C., Patterson, R. J., & Wolf, J. 2010, *ApJ*, 711, 671
- Kaplinghat, M., & Strigari, L. E. 2008, *ApJ*, 682, L93
- King, I. 1962, *AJ*, 67, 471
- King, I. R. 1966, *AJ*, 71, 64
- Kirby, E. N., Simon, J. D., Geha, M., Guhathakurta, P., & Frebel, A. 2008, *ApJ*, 685, L43

- Kirby, E. N., Lanfranchi, G. A., Simon, J. D., Cohen, J. G., & Guhathakurta, P. 2011, *ApJ*, 727, 78
- Kleyna, J. T., Geller, M. J., Kenyon, S. J., Kurtz, M. J., & Thorstensen, J. R. 1998, *AJ*, 115, 2359
- Kleyna, J., Wilkinson, M. I., Evans, N. W., Gilmore, G., & Frayn, C. 2002, *MNRAS*, 330, 792
- Kleyna, J. T., Wilkinson, M. I., Gilmore, G., & Evans, N. W. 2003, *ApJ*, 588, L21
- Kleyna, J. T., Wilkinson, M. I., Evans, N. W., & Gilmore, G. 2004, *MNRAS*, 354, L66
- Kleyna, J. T., Wilkinson, M. I., Evans, N. W., & Gilmore, G. 2005, *ApJ*, 630, L141
- Klimentowski, J., Łokas, E. L., Kazantzidis, S., Prada, F., Mayer, L., & Mamon, G. A. 2007, *MNRAS*, 378, 353
- Klypin, A., Kravtsov, A. V., Valenzuela, O., & Prada, F. 1999, *ApJ*, 522, 82
- Klypin, A., Kravtsov, A. V., Bullock, J. S., & Primack, J. R. 2001, *ApJ*, 554, 903
- Koch, A., Kleyna, J. T., Wilkinson, M. I., Grebel, E. K., Gilmore, G. F., Evans, N. W., Wyse, R. F. G., & Harbeck, D. R. 2007a, *AJ*, 134, 566
- Koch, A., Wilkinson, M. I., Kleyna, J. T., Gilmore, G. F., Grebel, E. K., Mackey, A. D., Evans, N. W., & Wyse, R. F. G. 2007b, *ApJ*, 657, 241
- Koch, A., Wilkinson, M. I., Kleyna, J. T., Irwin, M., Zucker, D. B., Belokurov, V., Gilmore, G. F., Fellhauer, M., & Evans, N. W. 2009, *ApJ*, 690, 453
- Koposov et al. 2008, *ApJ*, 686, 279
- Koposov, S. E., Yoo, J., Rix, H.-W., Weinberg, D. H., Macciò, A. V., & Escudé, J. M. 2009, *ApJ*, 696, 2179
- Koposov, S. E., Belokurov, V., Evans, N. W., Gilmore, G., Gieles, M., Irwin, M. J., Lewis, G. F., Niederste-Ostholt, M., Peñarrubia, J., Smith, M. C., Bizyaev, D., Malanushenko, E., Malanushenko, V., Schneider, D. P., & Wyse, R. F. G. 2011a, *ArXiv e-prints*
- Koposov, S. E., Gilmore, G., Walker, M. G., Belokurov, V., Wyn Evans, N., Fellhauer, M., Gieren, W., Geisler, D., Monaco, L., Norris, J. E., Okamoto, S., Peñarrubia, J., Wilkinson, M., Wyse, R. F. G., & Zucker, D. B. 2011b, *ApJ*, 736, 146
- Kormendy, J. 1985, *ApJ*, 295, 73
- Kormendy, J., & Freeman, K. C. 2004, in *IAU Symposium, Dark Matter in Galaxies*, Vol. 220, ed. S. Ryder, D. Pisano, M. Walker, & K. Freeman (San Francisco: Astronomical Society of the Pacific), 377-+
- Kravtsov, A. 2010, *Advances in Astronomy*, 2010
- Kroupa, P. 1997, *New Astron.*, 2, 139
- Kuhlen, M. 2010, *Advances in Astronomy*, 2010, 1-15
- Kuhn, J. R. 1993, *ApJ*, 409, L13
- Kuhn, J. R., & Miller, R. H. 1989, *ApJ*, 341, L41
- Kusenko, A. 2006, *Phys. Rev. Lett.*, 97, 241301
- Kuzio de Naray, R., McGaugh, S. S., de Blok, W. J. G., & Bosma, A. 2006, *ApJS*, 165, 461
- Kuzio de Naray, R., McGaugh, S. S., & de Blok, W. J. G. 2008, *ApJ*, 676, 920
- Lake, G. 1989, *AJ*, 98, 1253
- Law, D. R., & Majewski, S. R. 2010, *ApJ*, 714, 229
- Lee, M. G., Park, H. S., Park, J.-H., Sohn, Y.-J., Oh, S. J., Yuk, I.-S., Rey, S.-C., Lee, S.-G., Lee, Y.-W., Kim, H.-I., Han, W., Park, W.-K., Lee, J. H., Jeon, Y.-B., & Kim, S. C. 2003, *AJ*, 126, 2840
- Li, Y.-S., De Lucia, G., & Helmi, A. 2010, *MNRAS*, 401, 2036
- Lin, D. N. C., & Faber, S. M. 1983, *ApJ*, 266, L21
- Loeb, A., & Weiner, N. 2011, *Phys. Rev. Lett.*, 106, 171302
- Loewenstein, M., & Kusenko, A. 2010, *ApJ*, 714, 652
- Loewenstein, M., & Kusenko, A. 2012, *ArXiv:1203.5229*
- Loewenstein, M., Kusenko, A., & Biermann, P. L. 2009, *ApJ*, 700, 426
- Lovell, M. R., Eke, V., Frenk, C. S., Gao, L., Jenkins, A., Theuns, T., Wang, J., White, S. D. M., Boyarsky, A., & Ruchayskiy, O. 2012, *MNRAS*, 420, 2318
- Łokas, E. L. 2009, *MNRAS*, 394, L102
- Lynden-Bell, D. 1967, *MNRAS*, 136, 101
- Macciò, A. V., & Fontanot, F. 2010, *MNRAS*, 404, L16
- Macciò, A. V., Kang, X., Fontanot, F., Somerville, R. S., Koposov, S., & Monaco, P. 2010, *MNRAS*, 402, 1995
- Majewski, S. R., Ostheimer, J. C., Kunkel, W. E., & Patterson, R. J. 2000, *AJ*, 120, 2550
- Majewski, S. R., Skrutskie, M. F., Weinberg, M. D., & Ostheimer, J. C. 2003, *ApJ*, 599, 1082
- Majewski, S. R., Frinchaboy, P. M., Kunkel, W. E., Link, R., Muñoz, R. R., Ostheimer, J. C., Palma, C., Patterson, R. J., & Geisler, D. 2005, *AJ*, 130, 2677
- Majewski, S. R., Beaton, R. L., Patterson, R. J., Kalirai, J. S., Geha, M. C., Muñoz, R. R., Seigar, M. S., Guhathakurta, P., Gilbert, K. M., Rich, R. M., Bullock, J. S., & Reitzel, D. B. 2007, *ApJ*, 670, L9
- Mamon, G. A., & Łokas, E. L. 2005, *MNRAS*, 363, 705
- Martin, N. F., Irwin, M. J., Ibata, R. A., Conn, B. C., Lewis, G. F., Bellazzini, M., Chapman, S., & Tanvir, N. 2006, *MNRAS*, 367, L69
- Martin, N. F., Ibata, R. A., Chapman, S. C., Irwin, M., & Lewis, G. F. 2007, *MNRAS*, 380, 281
- Martin, N. F., de Jong, J. T. A., & Rix, H.-W. 2008, *ApJ*, 684, 1075
- Martin, N. F., McConnachie, A. W., Irwin, M., Widrow, L. M., Ferguson, A. M. N., Ibata, R. A., Dubinski, J., Babul, A., Chapman, S., Fardal, M.,

- Lewis, G. F., Navarro, J., & Rich, R. M. 2009, *ApJ*, 705, 758
- Martin, N. F., & Jin, S. 2010, *ApJ*, 721, 1333
- Martinez, G. D., Bullock, J. S., Kaplinghat, M., Strigari, L. E., & Trotta, R. 2009, *JCAP*, 6, 14
- Martinez, G. D., Minor, Q. E., Bullock, J., Kaplinghat, M., Simon, J. D., & Geha, M. 2011, *ApJ*, 738, 55
- Mashchenko, S., Couchman, H. M. P., & Wadsley, J. 2006, *Nature*, 442, 539
- Mashchenko, S., Wadsley, J., & Couchman, H. M. P. 2008, *Science*, 319, 174
- Mateo, M. L. 1998, *ARA&A*, 36, 435
- Mateo, M., Olszewski, E., Welch, D. L., Fischer, P., & Kunkel, W. 1991, *AJ*, 102, 914
- Mateo, M., Olszewski, E. W., Pryor, C., Welch, D. L., & Fischer, P. 1993, *AJ*, 105, 510
- Mateo, M., Mirabal, N., Udalski, A., Szymanski, M., Kaluzny, J., Kubiak, M., Krzemiński, W., & Stanek, K. Z. 1996, *ApJ*, 458, L13+
- Mateo, M., Olszewski, E. W., Vogt, S. S., & Keane, M. J. 1998, *AJ*, 116, 2315
- Mateo, M., Olszewski, E. W., & Walker, M. G. 2008, *ApJ*, 675, 201
- Mayer, L., Governato, F., Colpi, M., Moore, B., Quinn, T., Wadsley, J., Stadel, J., & Lake, G. 2001a, *ApJ*, 559, 754
- Mayer, L., Governato, F., Colpi, M., Moore, B., Quinn, T., Wadsley, J., Stadel, J., & Lake, G. 2001b, *ApJ*, 547, L123
- McConnachie, A. W., & Côté, P. 2010, *ApJ*, 722, L209
- McConnachie, A. W., Huxor, A., Martin, N. F., Irwin, M. J., Chapman, S. C., Fahlman, G., Ferguson, A. M. N., Ibata, R. A., Lewis, G. F., Richer, H., & Tanvir, N. R. 2008, *ApJ*, 688, 1009
- McConnachie, A. W., Irwin, M. J., Ibata, R. A., Dubinski, J., Widrow, L. M., Martin, N. F., Côté, P., Dotter, A. L., Navarro, J. F., Ferguson, A. M. N., Puzia, T. H., Lewis, G. F., Babul, A., Barmby, P., Bienaymé, O., Chapman, S. C., Cockcroft, R., Collins, M. L. M., Fardal, M. A., Harris, W. E., Huxor, A., Mackey, A. D., Peñarrubia, J., Rich, R. M., Richer, H. B., Siebert, A., Tanvir, N., Valls-Gabaud, D., & Venn, K. A. 2009, *Nature*, 461, 66
- McGaugh, S. S., de Blok, W. J. G., Schombert, J. M., Kuzio de Naray, R., & Kim, J. H. 2007, *ApJ*, 659, 149
- McGaugh, S. S., Rubin, V. C., & de Blok, W. J. G. 2001, *AJ*, 122, 2381
- Metz, M., & Kroupa, P. 2007, *MNRAS*, 376, 387
- Michie, R. W. 1963, *MNRAS*, 125, 127
- Milgrom, M. 1983, *ApJ*, 270, 365
- Minor, Q. E., Martinez, G., Bullock, J., Kaplinghat, M., & Trainor, R. 2010, *ApJ*, 721, 1142
- Moffat, J. W. 2006, *JCAP*, 3, 4
- Moore, B. 1994, *Nature*, 370, 629
- Moore, B., Governato, F., Quinn, T., Stadel, J., & Lake, G. 1998, *ApJ*, 499, L5+
- Muñoz et al. 2005, *ApJ*, 631, L137
- Muñoz, R. R., Carlin, J. L., Frinchaboy, P. M., Nidever, D. L., Majewski, S. R., & Patterson, R. J. 2006a, *ApJ*, 650, L51
- Muñoz et al. 2006b, *ApJ*, 649, 201
- Muñoz, R. R., Majewski, S. R., & Johnston, K. V. 2008, *ApJ*, 679, 346
- Muñoz, R. R., Geha, M., & Willman, B. 2010, *AJ*, 140, 138
- Muñoz, R. R., Padmanabhan, N., & Geha, M. 2011, *ArXiv:1110.1086*
- Navarro, J. F., Eke, V. R., & Frenk, C. S. 1996a, *MNRAS*, 283, L72
- Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996b, *ApJ*, 462, 563
- Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, *ApJ*, 490, 493
- Niederste-Ostholt, M., Belokurov, V., Evans, N. W., Gilmore, G., Wyse, R. F. G., & Norris, J. E. 2009, *MNRAS*, 398, 1771
- Norris, J. E., Wyse, R. F. G., Gilmore, G., Yong, D., Frebel, A., Wilkinson, M. I., Belokurov, V., & Zucker, D. B. 2010, *ApJ*, 723, 1632
- Odenkirchen, M., Grebel, E. K., Harbeck, D., Dehnen, W., Rix, H.-W., Newberg, H. J., Yanny, B., Holtzman, J., Brinkmann, J., Chen, B., Csabai, I., Hayes, J. J. E., Hennessy, G., Hindsley, R. B., Ivezić, Ž., Kinney, E. K., Kleinman, S. J., Long, D., Lupton, R. H., Neilsen, E. H., Nitta, A., Snedden, S. A., & York, D. G. 2001, *AJ*, 122, 2538
- Oh, K. S., Lin, D. N. C., & Aarseth, S. J. 1995, *ApJ*, 442, 142
- Okamoto, S., Arimoto, N., Yamada, Y., & Onodera, M. 2012, *ApJ*, 744, 96
- Olszewski, E. W., & Aaronson, M. 1985, *AJ*, 90, 2221
- Olszewski, E. W., Aaronson, M., & Hill, J. M. 1995, *AJ*, 110, 2120
- Olszewski, E. W., Pryor, C., & Armandroff, T. E. 1996, *AJ*, 111, 750
- Ostriker, J. P., Peebles, P. J. E., & Yahil, A. 1974, *ApJ*, 193, L1
- Pal, P. B., & Wolfenstein, L. 1982, *Phys. Rev. D*, 25, 766
- Palma, C., Majewski, S. R., Siegel, M. H., Patterson, R. J., Ostheimer, J. C., & Link, R. 2003, *AJ*, 125, 1352
- Parry, O. H., Eke, V. R., Frenk, C. S., & Okamoto, T. 2011, *ArXiv:1105.3474*
- Peñarrubia, J., Navarro, J. F., & McConnachie, A. W. 2008, *ApJ*, 673, 226
- Peñarrubia, J., Navarro, J. F., McConnachie, A. W., & Martin, N. F. 2009, *ApJ*, 698, 222

- Peñarrubia, J., Belokurov, V., Evans, N. W., Martínez-Delgado, D., Gilmore, G., Irwin, M., Niederste-Ostholt, M., & Zucker, D. B. 2010, *MNRAS*, 408, L26
- Piatek, S., & Pryor, C. 1995, *AJ*, 109, 1071
- Piatek, S., Pryor, C., Olszewski, E. W., Harris, H. C., Mateo, M., Minniti, D., Monet, D. G., Morrison, H., & Tinney, C. G. 2002, *AJ*, 124, 3198
- Piatek, S., Pryor, C., Olszewski, E. W., Harris, H. C., Mateo, M., Minniti, D., & Tinney, C. G. 2003, *AJ*, 126, 2346
- Piatek, S., Pryor, C., Bristow, P., Olszewski, E. W., Harris, H. C., Mateo, M., Minniti, D., & Tinney, C. G. 2006, *AJ*, 131, 1445
- Piatek, S., Pryor, C., Bristow, P., Olszewski, E. W., Harris, H. C., Mateo, M., Minniti, D., & Tinney, C. G. 2007, *AJ*, 133, 818
- Pieri, L., Lattanzi, M., & Silk, J. 2009, *MNRAS*, 399, 2033
- Plummer, H. C. 1911, *MNRAS*, 71, 460
- Polisenksy, E., & Ricotti, M. 2011, *Phys. Rev. D*, 83, 043506
- Pontzen, A., & Governato, F. 2011, *ArXiv:1106.0499*
- Pryor, C., & Kormendy, J. 1990, *AJ*, 100, 127
- Queloz, D., Dubath, P., & Pasquini, L. 1995, *A&A*, 300, 31
- Read, J. I., & Gilmore, G. 2005, *MNRAS*, 356, 107
- Read, J. I., Wilkinson, M. I., Evans, N. W., Gilmore, G., & Kley, J. T. 2006a, *MNRAS*, 367, 387
- Read, J. I., Wilkinson, M. I., Evans, N. W., Gilmore, G., & Kley, J. T. 2006b, *MNRAS*, 366, 429
- Richardson, J. C., Irwin, M. J., McConnachie, A. W., Martin, N. F., Dotter, A. L., Ferguson, A. M. N., Ibata, R. A., Chapman, S. C., Lewis, G. F., Tanvir, N. R., & Rich, R. M. 2011, *ApJ*, 732, 76
- Richstone, D. O., & Tremaine, S. 1986, *AJ*, 92, 72
- Romano-Díaz, E., Shlosman, I., Heller, C., & Hoffman, Y. 2009, *ApJ*, 702, 1250
- Salucci, P., & Burkert, A. 2000, *ApJ*, 537, L9
- Salucci, P., Wilkinson, M. I., Walker, M. G., Gilmore, G. F., Grebel, E. K., Koch, A., Frigerio Martins, C., & Wyse, R. F. G. 2011, *ArXiv:1111.1165*
- Sánchez-Salcedo, F. J., Reyes-Iturbide, J., & Hernandez, X. 2006, *MNRAS*, 370, 1829
- Sand, D. J., Strader, J., Willman, B., Zaritsky, D., McLeod, B., Caldwell, N., Seth, A., & Olszewski, E. 2011, *ArXiv:1111.6608*
- Saviane, I., Held, E. V., & Bertelli, G. 2000, *A&A*, 355, 56
- Sawala, T., Scannapieco, C., Maio, U., & White, S. 2010, *MNRAS*, 402, 1599
- Schwarzschild, M. 1979, *ApJ*, 232, 236
- Schweitzer, A. E., Cudworth, K. M., Majewski, S. R., & Suntzeff, N. B. 1995, *AJ*, 110, 2747
- Seitzer, P., & Frogel, J. A. 1985, *AJ*, 90, 1796
- Shapley, H. 1938a, *Harv. Coll. Obs. Bull.*, 908, 1
- Shapley, H. 1938b, *Nature*, 142, 715
- Simon, J. D., & Geha, M. 2007, *ApJ*, 670, 313
- Simon, J. D., Bolatto, A. D., Leroy, A., Blitz, L., & Gates, E. L. 2005, *ApJ*, 621, 757
- Simon, J. D., Geha, M., Minor, Q. E., Martinez, G. D., Kirby, E. N., Bullock, J. S., Kaplinghat, M., Strigari, L. E., Willman, B., Choi, P. I., Tollerud, E. J., & Wolf, J. 2011, *ApJ*, 733, 46
- Slater, C. T., Bell, E. F., & Martin, N. F. 2011, *ApJ*, 742, L14
- Sohn, S. T., Majewski, S. R., Muñoz, R. R., Kunkel, W. E., Johnston, K. V., Ostheimer, J. C., Guhathakurta, P., Patterson, R. J., Siegel, M. H., & Cooper, M. C. 2007, *ApJ*, 663, 960
- Spergel, D. N., & Steinhardt, P. J. 2000, *Phys. Rev. Lett.*, 84, 3760
- Spergel, D. N., Verde, L., Peiris, H. V., Komatsu, E., Nolte, M. R., Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Tucker, G. S., Weiland, J. L., Wollack, E., & Wright, E. L. 2003, *ApJS*, 148, 175
- Springel, V., Wang, J., Vogelsberger, M., Ludlow, A., Jenkins, A., Helmi, A., Navarro, J. F., Frenk, C. S., & White, S. D. M. 2008, *MNRAS*, 391, 1685
- Stecker, F. W. 1978, *ApJ*, 223, 1032
- Stetson, P. B., Hesser, J. E., & Smecker-Hane, T. A. 1998, *PASP*, 110, 533
- Strigari, L. E. 2010, *Advances in Astronomy*, 2010, 1–11
- Strigari, L. E., Bullock, J. S., Kaplinghat, M., Kravtsov, A. V., Gnedin, O. Y., Abazajian, K., & Klypin, A. A. 2006, *ApJ*, 652, 306
- Strigari, L. E., Bullock, J. S., Kaplinghat, M., Diemand, J., Kuhlen, M., & Madau, P. 2007a, *ApJ*, 669, 676
- Strigari, L. E., Koushiappas, S. M., Bullock, J. S., & Kaplinghat, M. 2007b, *Phys. Rev. D*, 75, 083526
- Strigari, L. E., Bullock, J. S., Kaplinghat, M., Simon, J. D., Geha, M., Willman, B., & Walker, M. G. 2008a, *Nature*, 454, 1096
- Strigari, L. E., Koushiappas, S. M., Bullock, J. S., Kaplinghat, M., Simon, J. D., Geha, M., & Willman, B. 2008b, *ApJ*, 678, 614
- Suntzeff, N. B., Aaronson, M., Olszewski, E. W., & Cook, K. H. 1986, *AJ*, 91, 1091
- Suntzeff, N. B., Mateo, M., Terndrup, D. M., Olszewski, E. W., Geisler, D., & Weller, W. 1993, *ApJ*, 418, 208
- The VERITAS collaboration: Vivier et al. 2011, *ArXiv:1110.6615*
- Tollerud, E. J., Bullock, J. S., Strigari, L. E., & Willman, B. 2008, *ApJ*, 688, 277
- Tollerud, E. J., Beaton, R. L., Geha, M. C., Bullock, J. S., Guhathakurta, P., Kalirai, J. S.,

- Majewski, S. R., Kirby, E. N., Gilbert, K. M., Yniguez, B., Patterson, R. J., Ostheimer, J. C., & Choudhury, A. 2011, ArXiv:1112.1067
- Tolstoy et al. 2004, ApJ, 617, L119
- Tolstoy, E., Hill, V., & Tosi, M. 2009, ARA&A, 47, 371
- Tonini, C., Lapi, A., & Salucci, P. 2006, ApJ, 649, 591
- Tremaine, S., & Gunn, J. E. 1979, Phys. Rev. Lett., 42, 407
- van den Bergh, S. 1972, ApJ, 171, L31
- van der Marel, R. P. 1994, MNRAS, 270, 271
- Vogelsberger, M., Zavala, J., & Loeb, A. 2012, ArXiv:1201.5892
- Vogt, S. S., Mateo, M., Olszewski, E. W., & Keane, M. J. 1995, AJ, 109, 151
- von Hoerner, S. 1957, ApJ, 125, 451
- Walcher, C. J., Fried, J. W., Burkert, A., & Klessen, R. S. 2003, A&A, 406, 847
- Walker, M. G., & Peñarrubia, J. 2011, ApJ, 742, 20
- Walker, M. G., Mateo, M., Olszewski, E. W., Bernstein, R., Wang, X., & Woodroofe, M. 2006, AJ, 131, 2114
- Walker, M. G., Mateo, M., Olszewski, E. W., Bernstein, R., Sen, B., & Woodroofe, M. 2007a, ApJS, 171, 389
- Walker, M. G., Mateo, M., Olszewski, E. W., Gnedin, O. Y., Wang, X., Sen, B., & Woodroofe, M. 2007b, ApJ, 667, L53
- Walker, M. G., Mateo, M., & Olszewski, E. W. 2008, ApJ, 688, L75
- Walker, M. G., Belokurov, V., Evans, N. W., Irwin, M. J., Mateo, M., Olszewski, E. W., & Gilmore, G. 2009a, ApJ, 694, L144
- Walker, M. G., Mateo, M., Olszewski, E. W., Peñarrubia, J., Wyn Evans, N., & Gilmore, G. 2009b, ApJ, 704, 1274
- Walker, M. G., Mateo, M., & Olszewski, E. W. 2009c, AJ, 137, 3100
- Walker, M. G., McGaugh, S. S., Mateo, M., Olszewski, E. W., & Kuzio de Naray, R. 2010, ApJ, 717, L87
- Walsh, S. M., Jerjen, H., & Willman, B. 2007, ApJ, 662, L83
- Walsh, S., Willman, B., & Jerjen, H. 2008, ArXiv:0807.3345
- Wang, X., Woodroofe, M., Walker, M. G., Mateo, M., & Olszewski, E. 2005, ApJ, 626, 145
- Watkins, L. L., Evans, N. W., Belokurov, V., Smith, M. C., Hewett, P. C., Bramich, D. M., Gilmore, G. F., Irwin, M. J., Vidrih, S., Wyrzykowski, Ł., & Zucker, D. B. 2009, MNRAS, 398, 1757
- Westfall, K. B., Majewski, S. R., Ostheimer, J. C., Frinchaboy, P. M., Kunkel, W. E., Patterson, R. J., & Link, R. 2006, AJ, 131, 375
- Wilkinson, M. I., Kleya, J., Evans, N. W., & Gilmore, G. 2002, MNRAS, 330, 778
- Wilkinson, M. I., Kleya, J. T., Evans, N. W., Gilmore, G. F., Irwin, M. J., & Grebel, E. K. 2004, ApJ, 611, L21
- Willman et al. 2005a, AJ, 129, 2692
- Willman et al. 2005b, ApJ, 626, L85
- Willman, B., Geha, M., Strader, J., Strigari, L. E., Simon, J. D., Kirby, E., Ho, N., & Warren, A. 2011, AJ, 142, 128
- Wilson, A. G. 1955, PASP, 67, 27
- Wolf, J., Martinez, G. D., Bullock, J. S., Kaplinghat, M., Geha, M., Muñoz, R. R., Simon, J. D., & Avedo, F. F. 2010, MNRAS, 406, 1220
- Zhao, H. 1996, MNRAS, 278, 488
- Zucker, D. B., Kniazev, A. Y., Bell, E. F., Martínez-Delgado, D., Grebel, E. K., Rix, H.-W., Rockosi, C. M., Holtzman, J. A., Walterbos, R. A. M., Annis, J., York, D. G., Ivezić, Ž., Brinkmann, J., Brewington, H., Harvanek, M., Hennessy, G., Kleinman, S. J., Krzesinski, J., Long, D., Newman, P. R., Nitta, A., & Snedden, S. A. 2004, ApJ, 612, L121
- Zucker, D. B., Kniazev, A. Y., Martínez-Delgado, D., Bell, E. F., Rix, H.-W., Grebel, E. K., Holtzman, J. A., Walterbos, R. A. M., Rockosi, C. M., York, D. G., Barentine, J. C., Brewington, H., Brinkmann, J., Harvanek, M., Kleinman, S. J., Krzesinski, J., Long, D., Neilsen, Jr., E. H., Nitta, A., & Snedden, S. A. 2007, ApJ, 659, L21
- Zucker et al. 2006a, ApJ, 650, L41
- Zucker et al. 2006b, ApJ, 643, L103

21 History of Dark Matter in Galaxies

Virginia Trimble

Department of Physics and Astronomy, University of California,
Irvine, CA, USA

1	<i>Introduction: Paleomatter and the Nature of History of Science</i>	1093
1.1	Paleomatter	1093
1.2	Some Characteristics of History of Science	1094
2	<i>When Stellar Statistics Was the Astronomical Forefront</i>	1095
2.1	The Stars in Motion	1095
2.2	Mapping the Milky Way	1096
3	<i>Dark Matter Between the Wars: Some Pairs of Papers</i>	1097
3.1	The Galactic Disk: Kapteyn and Jeans	1098
3.2	The Two Most Cited Early Papers: Oort and Zwicky	1098
3.3	Two Intermediate Scales: Holmberg and Babcock	1100
3.4	Interwar Summary	1101
4	<i>Interlude: 1939–1961</i>	1102
4.1	The Milky Way	1102
4.2	Other Individual Galaxies	1103
4.3	Binary Galaxies and Small Groups	1103
4.4	Rich Clusters	1104
4.5	The Universe as a Whole	1104
5	<i>1961–1974: From a Pair of Conferences to a Pair of Short Papers</i>	1104
5.1	The 1961 Conferences	1105
5.2	X-ray Considerations	1106
5.3	Nonstandard Gravitations and Cosmologies	1106
5.4	Gravitational Lensing (and Deflection of Light)	1107
5.5	The Cosmic Microwave Background and Its Remarkable Isotropy	1108
5.6	The First Non-Baryonic Candidates	1109
5.7	Rotation Curves Again	1109
5.8	Pure Theory: Disk Instabilities	1110
5.9	Big Bang Nucleosynthesis	1110
5.10	The 1974 Papers	1111

6	<i>Dark Matter and Standard Models Since 1974</i>	1111
6.1	The Impact of Inflation	1112
6.2	The Final Resuscitation of Λ	1112
6.3	Candidates: Dark, Darker, and Darkest	1113
6.4	Galaxy and Cluster Formation with Dark Matter	1113
6.5	Residual Doubts	1114
6.6	Dark Matter Alternatives	1115
6.7	From the Standard Hot Big Bang to Consensus Cosmology	1115
	<i>Acknowledgments</i>	1116
	<i>References</i>	1116

Abstract: The phrase dark matter goes back to 1922, and the concept of things that exist but are not, or cannot be, seen to the ancients. Here is one version of how the phrase was gradually restricted to kinds of matter not capable of electromagnetic interactions, how evidence for its existence very gradually accumulated, and how a number of watershed events have brought the astronomical community to near consensus that there is dark matter on many scales, but that its nature remains uncertain. The acceptance of dark matter in the universe has been described as a paradigm shift but also has many of the characteristics of normal science, in which data are acquired in response to existing ideas and those ideas gradually modified in light of the data. The approach is largely chronological, but with frequent looks ahead to see how various parts of the story turn out. It is not claimed that present understanding is complete or final.

Keywords: Cosmology, Dark matter, History of astronomy

1 Introduction: Paleomatter and the Nature of History of Science

That what we get is sometimes more than what we see is a very old idea in Western culture. The counter-earth of the Pythagorean Philolaus (fourth century BCE) was postulated to be forever undetectable. Kragh (2007, p. 213) regards it as the first example of dark matter in the history of cosmology. And the Latin church Credo speaks of all things seen and unseen, *visibilium et invisibilium*. Copernicus and Newton appear to have had no use for the *invisibilium*, but the idea was at least available for discussion. The decades from 1785 to 1845 brought forth a pair of ideas (black holes and dark stars) and a pair of observations (white dwarfs and Neptune) that are direct ancestors of some recent dark matter candidates. The year 1800 also saw the first deliberate search for something that might previously have just been too faint to be noticed. Only one of the seven items alluded to so far (entities with escape velocity larger than the speed of light) meets the present definition of dark matter: stuff that neither emits nor absorbs its fair share of radiation at any wavelength, but can interact with it gravitationally and perhaps weakly.

1.1 Paleomatter

The first idea was “unenlightened stars,” which might orbit and eclipse visible stars, producing periodic variability, first seen in Algol. Published by Edward Pigott (1805), the idea had been discussed between him and John Goodricke back around 1786 (Hoskin 1997, p. 202). The second idea now carries the name “black holes,” a phrase which first appears in print in *Science News Letters* for 19 January, 1964 (p. 39), but which gained popularity only later when used by John A. Wheeler in talks and papers (1968). The concept was phrased by John Michell (1784) as “all light emitted by such a body would be made to return to it by its own proper gravity,” and by Pierre Simon de Laplace (1796) as “il est donc possible que les plus grands corps lumineux de l’univers soient par cela meme invisible.”

The pair of observations occupied a number of years, but the analyses and announcements were nearly simultaneous (at least in our rest frame) in the 1840s. Friedrich W. Bessel (1845a, b) concluded that the paths of Sirius and Procyon across the sky were not straight, but oscillatory, so that each must have an invisible companion, and Urban J. J. Leverrier (1847, 1848) decided that irregularities in the motion of Uranus around the sun implied the existence of a more distant planet. John Couch Adams (1847) had reached a similar conclusion in the same time frame

as Leverier, but it was the latter's work that led to the 1846 discovery of Neptune by Galle at Berlin. The companions of Sirius and Procyon were discovered in 1862 (accidentally for Sirius by Alvan Clark) and in 1896 (after deliberate looking around Procyon by Schaeberle 1896).

In the meantime, F.X. von Zach (Hoskin pp. 188–190) attempted in 1800 to organize a team of 24 “celestial police” to scour the skies for the planet that, according to Bode's law, must come between Mars and Jupiter. They were scooped by Piazzi's accidental discovery of Ceres. He had not yet been told that he was supposed to be one of the policemen, though not for that zone. This is not untypical of much of the history of dark matter and, indeed, astronomy in general.

1.2 Some Characteristics of History of Science

First, it can never be complete. So far we have touched on seven possible entities and concepts – counter-earth, the invisibilium, unenlightened stars, black holes, white dwarfs (like Sirius B), gas giants (like Neptune), and small chunks of solid baryons (like Ceres and her smaller sisters found later), without having yet reached the twentieth century. They do not seem to have been brought together in any of many previous discussions of dark matter history, though all but the first have been suggested as DM candidates for recent decades. Kragh (2007) notes Philolaus and black holes; Jaan Einasto (2001) begins in 1915, and van den Bergh (2001) in 1933. My own previous ventures into this territory (Trimble 1987, 1988, 1995, 2005) also pick out only subsets.

Second, comparing the isotemporal articles by Einasto and by van den Bergh leads to the conclusions that claiming anything as “the first” is very dangerous, and that history can look very different to folks on opposite sides of borders (the Oder-Neiss line in this case).

In addition, the professional historians themselves are not in full agreement about how one ought look at things – in any case, not just “who got it right first” (which is called Whiggish) or “who persuade the community” (populist), but perhaps more nearly, “why did X think Y, who influenced him, whom did he influence, and how?” The approved perspective changes with time. I think the concept of a paradigm shift (Thomas Kuhn in 1962) had already gone out of fashion by the time Tremaine (1987) described the recognition of dark matter in those terms. He put the beginning of the shift in 1974 and predicted that things should be sorted out by 1994 (perhaps just barely true, though I date the current standard model to 1997 at the Kyoto General Assembly of the International Astronomical Union).

Fourth and finally, progress is hardly ever monotonic. How did two such sensible ideas as eclipses by dark bodies and $2GM = Rc^2$ ever get dropped from the astronomer's tool kit? In the case of the stars, observations of additional ones (especially Delta Cephei) revealed variables with imperfect periodicity, non-symmetric light curves, and other irregularities that Keplerian orbits could not match, leaving the older idea of spotted, rotating stars (like the sun only more so) master of the terrain until the late nineteenth century. Binaries were then the best buy model for all variables until analysis of Delta Cephei and others showed that one star would have to be inside the other, leading to pulsation as a third mechanism. All three, of course, actually happen, occasionally in a single system.

As for the bodies that held back their own light, the problem was that $2GM/Rc^2 = 1$ had been calculated with a particle theory of light, Newtonian mechanics, and the assumption that the process would manifest itself as light being slowed down as it left massive sources. The wave theory of light took over in the early nineteenth century, leaving no obvious way to calculate the phenomenon for bodies with less extreme M/R . Observationally, the slowing of light would have shown up as color (or at least Doppler) shifts and as loss of synchronism of orbits of the

two stars in visual binaries. Nothing of the sort was seen. Thus the whole realm of interaction of light with gravitational fields had to be rethought in the light of special and general relativity, beginning with Einstein himself (deflection of light; gravitational redshift), Chwolson (gravitational lensing), Schwarzschild, and all the rest. That the classical and relativistic expressions for the Schwarzschild radius and (to first order) gravitational redshift are the same counts as one of the minor mercies of twentieth century astrophysics. Light bending, which had also been calculated in the particle/Newtonian era, is a factor of two larger in the relativistic case. This comes into the evaluation of alternatives to dark matter in their ability to describe the velocity distributions, X-ray temperatures, and gravitational lensing for clusters of galaxies.

Notice that, in contrast to the theoretical cases, the observed entities have never been out of the astronomical inventory. As for the “celestial police,” many astronomers since have tried to carry out major programs by persuading their colleagues to help out (the *Carte du Ciel* in the 1880s; Zwicky’s supernova search programs in the 1960s). There must be a better analogy than cat-herding, but it is really only in the last decade or two that large teams have come together in large survey or monitoring programs using many facilities, some purpose-built. The driver has been the need for a large constituency to pry loose large numbers of euros, dollars, and all. Recent dark matter searches have begun to approach this condition.

2 When Stellar Statistics Was the Astronomical Forefront

Each era has had its handful of astronomical questions generally regarded as important. Today’s seem largely to be in the territory of origins and formation – of the universe, of galaxies, of stars and planets – with dark matter (and dark energy) part of at least the first two. New observations and interpretations in each epoch will generally make sense only within the framework of those important questions. Once upon a time, from the Greeks to Tycho and Kepler, to Newton and Halley and beyond, it was obtaining accurate observations of the moon, sun, planets and their moons, and comets and fitting them with a theory that would permit prediction of eclipses, transits, conjunctions, and other mutual events and returns. Through most of this period, the “fixed stars,” apart from an occasional nova stella or variable, were merely the pattern against which important motions could be measured, though Newton was of the opinion that Divine Intervention might be necessary to keep them fixed for long periods (Kragh 2007, p. 73 on 1692–1697 correspondence with Nicholas Bentley).

2.1 The Stars in Motion

But by 1700, the stars were about to break loose from their moorings, in four steps. These, in chronological order of discovery (broadly interpreted!) are called proper motions, binary stars, parallax, and radial velocities. First, Edmund Halley (1718) announced that three of the brightest stars had coordinates different (above and beyond precession) from those reported by the ancients. James Bradley’s 1729 recognition of aberration of starlight is to us a triumph of precision measurement which required the earth to orbit the sun rather than conversely. But it was then also a major source of noise in efforts to determine accurate stellar positions and motions. The names associated with the gradual accumulation of meaningful proper motions (Hoskin 1997, pp. 202–209) were Bessel, Tobias Mayer (who pointed out in 1760 that it was

relative motion of the star and sun being seen), William Herschel, F.W.A. Argelander, and Thomas Galloway. All attempted to discern the motion of the sun relative to all the other stars; their answers were largely concordant (including that of Galloway, who used an independent southern sample) and in agreement with the modern solar apex, toward Hercules.

Bradley had, of course, been looking for parallax, and so was Herschel when he (1803) recognized that a few close pairs of stars in the sky were moving around each other. That there were too many close pairs and clusters to be chance superpositions had been noticed and published by Michell (1768) but largely forgotten. That the orbits showed Newtonian gravitation beyond the solar system became certain with the first eccentric pair (Xi UMa, charted by Felix Savary in 1827).

The long-sought parallax appeared in three sets of measurements of three different stars by two famous astronomers (Bessel yet again and Wilhelm Struve) and one rather obscure one (Thomas Henderson who had the advantage of working from the Cape of Good Hope and looking at Alpha Centauri) during 1835–1838. The values found, all less than 1 arc sec, confirmed the very large distances (10^6 AU or parsecs and more, though neither unit yet had that name) implied by apparent stellar brightnesses and the idea that stars were suns. Of course observing the sun and other stars at the same time presented certain difficulties! Huygens tried putting a screen with a pin-prick hole between himself and the sun.

But the hole was too big, and he ended up putting Sirius at only 27,664 AU. James Gregory in 1688 suggested using a planet as an intermediary and got 83,190 AU (Hoskin p. 211), and Newton did still better but did not publish until 1728. The modern value of course exceeds 200,000 AU, but apparent brightnesses and then parallaxes in any case made clear that the “universe of stars” was very large indeed compared to the solar system. Though the inventory of reliable parallaxes also grew very slowly, they enabled the conversion of proper motions to linear speeds and the use of visual binary orbits to measure a few stellar masses, which indeed proved comparable with that of the sun.

Radial velocities required the development of spectroscopy and so came last of the four. Huygens tried but failed (again), and the first handful of meaningful numbers – meaning errors of a few km/s for values of 10–30 km/s – came from H.C. Vogel and Julius Scheiner at Postdam and James Keeler at Lick around 1890. The first spectroscopic binary orbits followed soon and provided a few more stellar masses.

With the stars now in motion, it became possible to ask about the statistics of numbers vs. luminosity, distance, mass, proper motion, and radial velocity; to calculate the amount of mass necessary to account for the motions; and to ask whether the stars themselves added up to the necessary mass.

2.2 Mapping the Milky Way

It is in this context that astronomers first found some hint of dark matter in the Galaxy (or universe) in the early twentieth century. In the same time frame, determinations of the solar motion relative to, first, more distant stars, then globular clusters, and finally external galaxies led up to the velocity–distance relation which is called Hubble’s law.

None of the pre-1918 data strongly contradicted the widely held opinion, going back at least to Herschel, that the solar system was near the center of a thick disk of stars a few thousand light years (or parsecs, to within a factor three!) in diameter. Many of the best-known astronomers of the period worked on problems of stellar statistics, including Kapteyn, Jeans, Eddington, Karl

Schwarzschild, Strömberg, and later Oort. They invented descriptors like star streams, velocity ellipsoids, and asymmetric stellar motions. All would eventually be subsumed in the idea of galactic rotation (of the disk, but not the halo) and velocity dispersions around the average rotation in radial, angular, and perpendicular directions, R , θ , and Z .

An interesting snapshot of the community struggling with these various concepts, after the recognition by Shapley of the off-center position of the sun and by Hubble of the existence of other galaxies, but before the discovery of rotation, is found in Russell et al. (1926, Vol II), which was the primary astronomy text for a generation of our English-speaking predecessors. Russell et al. reproduce an argument due to Kelvin, which says that the volume occupied by a uniform density of stars like that near us cannot be either very small or arbitrarily large, or stars would either leave, in the small case, or have much larger speeds than the largest seen (about 350 km/s then) in the large case. Conversely, the largest speeds plus a Galaxy size of Herschelian or Kapteynian dimension require an average density of about one low mass star (the commonest sort) per cubic parsec. Russell et al. state specifically that the argument applied to the density of “all matter, whether luminous or dark.”

This was the context in which Ernst Opik (1915) set out to find the density of matter near the Galactic plane, using the speeds of stars and their distances from it. He concluded that the required mass might well all be in the stars themselves, assuming a mass-to-light ratio of 2.63 (in solar units) implied by counts of stars of various spectral types. He combined his result with the apparent surface brightness of the central part of the Andromeda Nebula, velocity information from Slipher’s (1914) spectrum, and Newton’s laws to deduce a distance of 440 kpc for M31 (Opik 1922), about half the current value and obviously well outside the Milky Way.

Notice that Opik was essentially right both about the minor importance of local dark matter and about the existence of external galaxies at a time when the majority of the astronomical community would have voted with writer and popularizer Agnes Mary Clerke, whose pair of books in 1896 and 1903 (Hockey et al. 2007, entry for A.M. Clerke) denied with great vehemence the existence of other observable galaxies but included a whole chapter called “Dark Stars.” By this she meant ones that had faded following exhaustion of their (unknown) energy supply. They are called white dwarfs, and they have been dark matter candidates in the past.

The story of stellar energy production is an almost completely separate one from the dynamical issues addressed here, but Simon Newcomb (1906), first president of the American Astronomical Society, described the source of solar and stellar energy as the most important unsolved astronomical problem of his time, though his own work was largely in positional astronomy. He also questioned the central position generally assigned to us in the Galaxy (by analogy with Ptolemaic cosmology) and asked himself whether the perihelion advance of Mercury might perhaps be attributable to a small deviation from Newtonian gravity.

3 Dark Matter Between the Wars: Some Pairs of Papers

Let us deal first with the divisive question of who coined the phrase dark matter. Zwicky (1933), whom you will meet properly in [Sect. 3.2](#), writing in German, spoke of “dunkle Materie” and is often given credit, but “dark matter” in English will be seen appearing a decade earlier. The name has taken a very long time to become fully standard. In 1977, Ivan King, Martin Rees, and James Gunn (separately, in Tinsley and Larson 1977) all wrote “missing mass” or “unseen mass. J.P. Ostriker enunciated at that meeting, before, and after it, that it is the light that is missing, not

the mass. King countered that the mass was in any case “missing from our understanding” and stuck by the old phrase. Dark matter was, however, the winner over the next few years (Trimble 1987). There have also been holdouts against the very notion, some of whom appear in later sections. If you are young enough that 1977–1987 is history for you, rather than current events, then you may even need reminding that the two wars were those of 1914–1918 and 1939–1945. And if someone should try to sell you a document said to have been written before 1940 which speaks of World War I, it is a forgery.

3.1 The Galactic Disk: Kapteyn and Jeans

In the years after 1915, the number of accurate radial velocities, stellar distances, and masses accumulated rapidly. Some of the work was driven by J.C. Kapteyn’s effort to map out the entire Milky Way, using his Selected Areas, and his (Kapteyn 1922) estimate of the mass in the Galactic disk is a fairly small part of that program. The 1922 paper was written only about a year before his death, and his decision to assume no interstellar absorption or scattering of starlight was perhaps driven by the feeling that he must finish the paper soon or not at all. Jeans (1922) also used velocities and distances perpendicular to the plane to calculate the mass density in it, without at that time attempting to map the entire Galaxy.

Their calculations of the local mass density are not invalidated by their assumption that the solar system was very near the center of the Milky Way, since only stars within about a kiloparsec were used. Their results (0.143 and $0.099 M_{\odot}/\text{pc}^3$) roughly bracket more recent values, as do their conclusions, based on comparing the mass per luminous star implied by their numbers with the average mass of known binary systems ($1.6 M_{\odot}$ at that time).

Jeans opined that “there must be about three dark stars in the universe for every bright star,” while Kapteyn, noting that “we have therefore the means of estimating the mass of dark matter in the universe,” concluded that “as matters stand at present, it appears at once that this mass cannot be excessive.” Their universe is, of course, our Galaxy, and was not capitalized.

The Jeans and Kapteyn conclusions are, more or less, phrased respectively as a detection of dark matter and as an upper limit. The two came gradually together, once faint stars, stellar remnants, and gas were promoted from dark to enlightened and measurements of stellar distances outside the plane became more numerous and accurate. We are, incidentally, less than 20 pc from the median plane.

The current status of dark matter in galactic disks is that they will have a slightly larger density than the adjacent halo because of dark matter brought in, along with baryons, in mergers and satellite captures (Read et al. 2006, 2008). Two short-lived alternatives were the possibility of very dense, compact clouds of molecular hydrogen in outer disks (ruled out by absence of absorption in them, Clarke et al. 2003) and outer light that had been missed (Valentijn 1990), eventually ruled out by deeper infrared imaging.

3.2 The Two Most Cited Early Papers: Oort and Zwicky

More data for more stars continued to appear, and what Oort (1932) brought to the table was a much more sophisticated way of analyzing the data, involving Poisson equations and so

forth. He had by then also provided from stellar motions strong evidence for galactic rotation (Oort 1927), which had been proposed the year before by Lindblad (1926). The most accessible discussion of the methods is probably Oort (1965), in which he describes the goal of stellar dynamics as finding relations between the density and velocity distributions of stars. The 1932 result for local mass density was 0.08 to 0.11 M_{\odot}/pc^3 , so that one solar luminosity corresponded to 1.8 solar masses. Stars brighter than $M_{\odot V} = 13.5$ contributed 0.038 M_{\odot}/pc^3 . Oort's opinion then was that stars between $M_V = 13.5$ and 18.5 would account for most of the rest. This was based heavily on apparent (and erroneous) gravitational redshifts near 240 km/s for the white dwarfs Procyon B and van Maanen 2, which (using R.H. Fowler's non-relativistic equations for degenerate matter) would have required them to have masses much larger than 1 M_{\odot} .

Oort's 1965 number was 0.148 M_{\odot}/pc^3 , of which about 40% had to be attributed to stars or gas of unknown type. (HI had been inventoried by that time, but not molecular hydrogen.) This number is still often called the Oort limit, though by now it can be seen that it should really be the Kapteyn–Jeans limit or even the Opik limit.

In addition to the local mass density, one would like to know M and M/L for a cylinder extending far out perpendicular to the disk, so that the numbers can be compared with those for other galaxies as well as telling us something about the nature of dark matter in our own. Oort (1965) pointed out that, if the unknown 40% were distributed like halo stars, then M/L in the cylinder would be as large as 13.4 (but only 2–6 for the 40% all in a thin plane distributed like old disk stars). He regarded the larger number as almost impossible, though it was comparable with what de Vaucouleurs (1958) had found for a similar cylinder about 10 kpc from the center of M31. It could now be said that these large numbers made sense if the cylinders extended to 100 kpc, through the galactic halo, but they did not, because there are no tracer stars of the disk population more than a couple of kpc out of the plane.

Before moving on to the other, Zwicky, paper, recall that, for anything dominated by gravitation, a reasonable mass estimate is $M = V^2 R/G$ for properly selected velocity, V , and length scale, R . V is mercifully distance-independent, but R (hence M) is, of course, linear in distance and, therefore, in the reciprocal of the Hubble constant H , for things outside the Local Group. Luminosity corresponding to measured apparent brightness goes as R^2 , so that M/L values for extragalactic entities are linear in H . Between the first 1929 Hubble paper and the Rome 1952 IAU General Assembly, nearly everyone took H to be close to 500 km/s/Mpc (Trimble 1996 mentions a few early exceptions). Inverted, this is the same arithmetic Opik (1922) used to estimate the distance to M31. Hubble's 1934 estimate of $10^9 M_{\odot}$ for a typical galaxy assumed M/L like the solar neighborhood (i.e., no dark matter) and included only central bright regions.

No recent astronomer seems to have doubted the existence of galaxies outside ours, but reservations about clusters persisted. Hubble thought they were rare, while Shapley thought truly isolated galaxies were rare. Partly they differed in definition: Hubble meant things like Coma and Virgo, while Shapley included the Local Group, which from 100 Mpc would look like a binary of the Milky Way and M31. And, thought Shapley, Hubble had access to excessively large telescopes, whose small fields of view made it hard to see the big picture. Indeed “cluster denial” has persisted almost to the present. For instance, Fasenko (1985) and Zabrenowski (1986) and other papers by the same authors attribute the patchy distribution of galaxies on the sky to differential absorption in the Milky Way, in which case measured velocity dispersions are merely random fluctuations around the Hubble flow.

The second now highly cited paper is that of Zwicky (1933), who realized that for a cluster of anything the relevant V is the velocity dispersion and turned the 100” telescope at Mt. Wilson

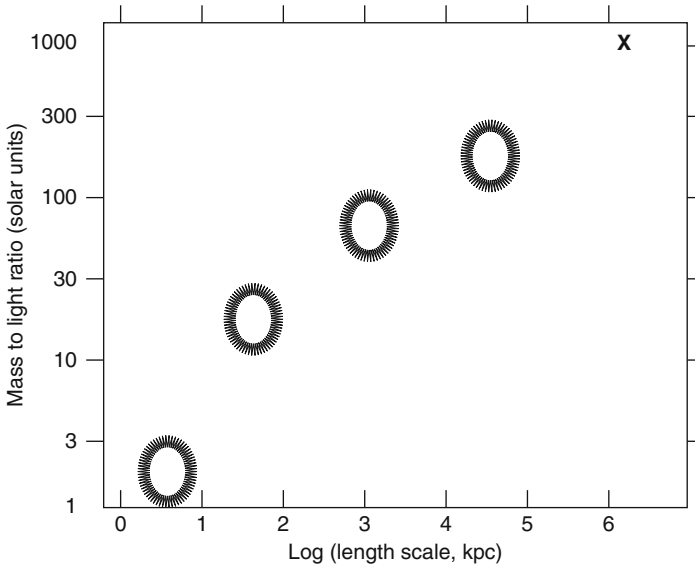
toward the Coma cluster. He measured eight radial velocities and found a dispersion of about 1,000 km/s, larger than he had expected, and leading to his conclusion “dass dunkle Materie in sehr goesserere Dichte vorhanden is als leuchtenden Materie.” That is, a thousand “Hubble” galaxies of $10^9 M_\odot$ each would add up to only $10^{12} M_\odot$, and he had found something more like $10^{14} M_\odot$. This number is not called the Zwicky limit even now, though he is generally now cited as a dark matter pioneer. Zwicky supposed that dwarf galaxies and gas must make up the missing 99%, later adding intergalactic stars and pygmy galaxies to his inventory. He looked for evidence of these as excess sky brightness between the known Coma galaxies (Zwicky 1951, 1958) and thought he had seen an appropriate number of photons. Intracluster light and indeed long-lived entities like globular clusters and planetary nebulae have been reported many times since, but they do not add up to more mass than is in the recognizable galaxies (Da Rocha and Mendes de Oliveira 2005).

A few years later, Smith (1936) used a spectrograph of his own design to make the Mt. Wilson 60” nearly as powerful as the 100” for some purposes. He aimed the combination toward the Virgo cluster and, from 32 radial velocities, found a total mass of $10^{14} M_\odot$ (for a distance of 2 Mpc, thus about $9 \times 10^{14} M_\odot$ at a modern distance). He noted that the “implied mass per nebula is $2 \times 10^{11} M_\odot$... far larger than Hubble’s value of $10^9 M_\odot$ for the mass of an average nebula.” He went on to say that “it is possible that both figures are correct and that the difference represents a great mass of internebular material within the cluster.” Indeed it does.

3.3 Two Intermediate Scales: Holmberg and Babcock

Holmberg’s (1937) thesis primarily addressed the demonstration that there must be bound pairs of galaxies, because there were too many close pairs on the sky to be chance superpositions, very much like Michell’s 1768 argument for binary stars, but Holmberg did a better job of the statistics. As a by-product of the measured pair separations on the sky and in radial velocity, he ended up with an estimate of average mass per galaxy or M/L for binaries. His average mass was $10^{11} M_\odot$ (perhaps $5 \times 10^{11} M_\odot$ on a modern distance scale). Remarkably, given the large errors everywhere, his number fit comfortably between Hubble’s single galaxies and Zwicky’s Coma cluster. He cited Zwicky (1933) and Smith (1936), as well as Hubble and Shapley, and stated that his value was a satisfactory one “in very good agreement with the values earlier derived.” He remained an outstanding extragalactic astronomer throughout his career.

Babcock (1939), who measured the first extended rotation curve for a spiral galaxy (M31 naturally), left the field and became primarily known for contributions to solar physics, especially the theory of the 22 year activity cycle. He explained why in a handwritten letter responding to my inquiry about why no-one (including himself) had followed up on his work for such a long time. His rotation curve was still rising at the last measured point and (scaling to a distance of 785 kpc) implied $M = 3 \times 10^{11} M_\odot$ and $M/L = 17$ out to 18 kpc. His paper remarked upon the difference from the Milky Way as then understood and did not cite Zwicky, Smith, or Holmberg. A public talk on his results (at the 1939 dedication of McDonald Observatory) was heavily criticized by other meeting participants, and he was instructed to publish his thesis as a Lick Observatory Bulletin rather than in the *Astrophysical Journal*. He found neither of these events encouraging, though curiously his contemporary, Daniel M. Popper, also at Lick was instructed to publish in *ApJ*, when he would have preferred the *LOB* series! Babcock’s last data point, 380 km/s 100’ from the center of M31 came from a multi-night exposure and



■ Fig. 21-1

Illustrative graph of M/L ratios vs. length scale, showing the increasing dominance of dark matter for larger structures, from galaxy cores to clusters. Ellipses indicate numbers that could have been plotted as early as 1939, but were not

was probably erroneous, though the resulting mass and M/L now sound right on. Vera Rubin (personal communication and talks at meetings) has wondered whether he might perhaps have had doubts about the numbers even at that time.

The second rotation curve ever measured was that of M33 (Mayall and Aller 1960). M33 is now known to be tidally distorted and truncated, so that the small mass and M/L they found were inevitable. It is hard to say in retrospect whether this was good or bad luck.

3.4 Interwar Summary

A widely read astronomer of trusting spirit could, in 1939, have plotted a little graph, like [Fig. 21-1](#), showing a monotonic, nearly linear increase of M/L with length scale (at least in log-log coordinates!). This was not done at the time and perhaps would not have suggested to 1939 eyes dark matter more widely distributed than the luminous. There was one version in tabular form due to Schwarzschild (1954) which he did not interpret as an M/L gradient, and such plots did not appear for another 20 years. Indeed none of the data were much discussed, and van den Bergh (2001) presents in some detail the extent to which the Zwicky (1933) and Babcock (1939) papers were ignored. Zwicky's (1937) suggestion that gravitational lensing by galaxies would firmly settle the issue of their masses belongs chronologically to this period but intellectually to the modern, dark-matter-dominated era.

4 Interlude: 1939–1961

Two fairly obvious things kept this generation from being a golden one of dark matter research. First, there was a war going on and then a good deal of essential postwar reconstruction, especially in Europe (from Anglesey to Kiev). Second it was a golden age of stellar structure and evolution research, because a basic understanding of nuclear reactions as the energy source for the sun and stars had been achieved just as WWII broke out and because computational power was rapidly expanding during and after the war. Thus, many outstanding astronomers focused on stars. As for our pioneers, Jeans (1877–1946) and Kapteyn (1851–1922) had died full of years and honors and Sinclair Smith (1899–1938) tragically early of cancer. Zwicky (1898–1974) and Oort (1900–1992) continued to advocate dark matter and many other things up essentially until the times of their deaths.

Nevertheless, data suggestive of large M/L ratios in various contexts gradually accumulated. The contexts included the Milky Way, other individual galaxies, pairs (including the Local Group), rich clusters, and the universe as a whole. There were also papers that found mass-to-light ratios fully accounted for by normal stars and gas, most notably a series by Margaret Burbidge and Geoffrey (1961 and references therein) on rotation of spiral galaxies, in which, however, they repeatedly emphasized that they could say nothing about faint outer regions where the photographic plate detectors of the time did not permit recording either emission or absorption line spectra. In case you had ever wondered, Slipher's spectrum of M31 showed absorption features from the stars most clearly and Babcock's the emission lines of HII regions.

The other distractor was a pair of papers on binary galaxies by Page (1952, 1956). He differed from Holmberg in requiring smaller separations on the sky and in radial velocity before declaring a pair to be physically bound. Smaller average mass per pair was an inevitable result.

Items are arranged by length scale rather than by year in the rest of this section.

4.1 The Milky Way

Calculated masses for our Galaxy interior to the solar circle depend on the adopted values of local rotation speed (V_c) and galactocentric distance (R_o), with detailed modeling of flat vs. spheroidal distributions less important. Thus this mass has shrunk from somewhat larger than $10^{11} M_\odot$ to somewhat smaller over the past few decades, as best estimates of V_c have declined from about 250 to about 220 km/s and of R_o from 10 to 8–8.5 kpc (Ghez et al. 2008). But it is the mass outside R_o that is important for dark matter, and this is much less certain. Something near $1 \times 10^{11} M_\odot$ for the inner Galaxy would have been fine with Oort (1932) and also Kuzmin (1952) and Schmidt (1956) among the early modelers.

The population of globular clusters extends considerably beyond the solar circle, and the dispersion of their velocities provides a handle on the larger scale mass. Knowing only the radial component is a challenge, but Kunth (1952) and Lohman (1956) each found two or three times as much mass outside R_o as inside, and more elaborate analyses confirmed their conclusion (Table 1 of Tremaine 1987), though with results spread across $0.3\text{--}30 \times 10^{11} M_\odot$. More recently, proper motions for some globular clusters have been measured, and the inventory of dwarf galaxy companions has expanded to the point where they can be used the same way. The conclusion that outer (halo) mass considerably exceeds inner (disk + bulge) mass holds.

4.2 Other Individual Galaxies

The third published rotation curve was Oort's (1940) for the S0 galaxy NGC 3115. He stated that "the distribution of mass in this object appears to bear almost no resemblance to that of light" (a view which the Schwarzschild 1954 paper was partly designed to refute). Van den Bergh (2001) attempts to understand Schwarzschild's view, but the explanation bears some of the signs of a cracking paradigm – large error bars, data that should be left out of the sample, and emphasizing data that agree with the predetermined conclusion, like the rotation curve of M33. Schwarzschild did accept that the visible parts of elliptical galaxies had larger M/L ratios than the corresponding parts of spirals. He attributed the difference to a large population of white dwarfs in the ellipticals. Oort's candidate for NGC 3115 was faint M dwarfs.

The advent of radio astronomy would eventually settle the issue for rotating galaxies, beginning with van de Hulst's (1957) confirmation of a global M/L near 20, vs. 2 in the nucleus (Lalleman 1960), of M31. There do, of course, also exist some rotating galaxies with rotation curves that turn over at large radii, (Cicaire and et al. 2008), and the case for dark matter in ellipticals was historically more difficult to make because one needed to use absorption lines that are the sums of many stars on sight lines through various parts of the galaxy, until radial velocities for their globular clusters, X-ray temperatures, and lensing data (Pooley et al. 2008), became available.

4.3 Binary Galaxies and Small Groups

Page (1960) reconsidered his sample, corrected some numerical errors, and endorsed Holmberg's (1937) larger M/L values. So, not so surprisingly, did van den Bergh (1961) and Holmberg (1954) himself, using larger samples and a less stringent criterion for which pairs to count as bound. At least six papers (cited in Trimble 1995, [▶ Sect. 4](#)) analyzed small groups and found mass-to-light ratios up to about 100.

The most innovative discussion was surely that of Kahn and Woltjer (1959), who brought together the ideas that our Galaxy and Andromeda are approaching each other (known to Slipher 1914) and that the universe is expanding (if not quite known to Hubble in 1929 then to others soon after). The actual approach speed is about 100 km/s, part of the observed 300 km/s being due to Milky Way rotation; and the Universe of 1959 was about 10 Gigayears old. In that time, the gravitating masses of the two galaxies must have been enough to stop their motion apart from each other and turn them around to start coming back together again. Not knowing the relative transverse speed or whether the upcoming close passage will be the first were obvious sources of uncertainty, but could not force the total mass far away from 10 times the numbers for the inner 10 kpc of each.

The authors supposed that the other 90% ought to be gas in some temperature/density regime not (yet) detectable. Oddly, in the proceedings of IAU Symposium 35, Woltjer is quoted as saying that this $10^{12} M_{\odot}$ of gas was "a number that had crept into the literature and now seems to be creeping out." Much of the gas, was, as it were, forced out by advances in radio and X-ray astronomy, but there is arguably still some supply at about 10^5 K that continues to fall in as high-velocity clouds and to impose ultraviolet absorption lines of multiply ionized oxygen (etc.) on the spectra of more distant sources. This belongs to the "missing baryon problem" (Richter et al. 2006).

4.4 Rich Clusters

Between 1950 and 1960, the Virgo and Coma clusters were reexamined at least three times each and the clusters in Hercules and Canum Venaticorum at least once each (references in Trimble 1995, [▶ Sect. 4](#)). As a rule, there were tens of radial (one-dimensional) velocities out of a thousand galaxies, no transverse velocity information, and the best value of Hubble's constant was declining rapidly from 500 to perhaps 125 km/s/Mpc (Trimble 1996). Nevertheless there was striking general agreement that cluster mass-to-light ratios ranged upwards from 100 to perhaps as much as 1,000. Recall that, for a cosmic luminosity density of $10^8 L_{\odot}/\text{Mpc}^3$ (another of those numbers that goes back to Oort and has not changed much), M/L a bit larger than 1,000 takes us to the critical, closure density for a standard relativistic universe.

4.5 The Universe as a Whole

Einstein's equations permit a very wide range of values for the Hubble constant, H , the density and pressure of stuff, ρ , and the cosmological constant. Early cosmologists (Tolman 1934, for instance) considered many possibilities including some with no epoch of very high density and oscillating models that cannot apply to our universe.

In contrast, the Steady State universe of Bondi and Gold (1948) and of Hoyle (1948) requires precisely the critical density at all times, presumably in hydrogen gas available to form new galaxies, at a temperature greater than 10^5 K (Hoyle 1959), which at the time could not have been ruled out observationally. This requirement for a very large amount of undetected matter was one (though not the only) source of violent objections to the model (Kragh 1996). Other alternatives to Einsteinian cosmology, e.g., those of Dirac and Milne, also required a critical density (Kragh 2007, [▶ Sect. 3.4](#)). Dirac held by such a model through most of his life, though the version he spoke about at a conference in the 1970s at a planetarium dedication would have given us a sky glowing with H-alpha emission that you could almost read by. This was also the meeting where, at luncheon, he asked his neighbor at the table, "Do you want my ice? Ice causes flu."

Despite the wide range of possible general relativistic universes, the choices $\Lambda = 0$ and $q_0 = 1/2$ (critical density) were very popular in mainstream texts and articles. There seem to have been three reasons for the choice: (1) those conditions could not be ruled out (unless you worried about ages of things), (2) $k=0$ (flat space) and $\Lambda = 0$ together have an aesthetically pleasing appearance, and (3) the calculations needed for analyzing data (apparent magnitudes and angular diameters vs. redshift, for instance) are thereby much simplified. Non-zero k 's and Λ 's, however, also appear scattered through the mainstream literature, motivated by considerations of the formation of helium in the early universe, ages of things vs. various values of the Hubble constant, and the difficulties of forming galaxies and clusters in an expanding universe.

5 1961–1974: From a Pair of Conferences to a Pair of Short Papers

At least half a dozen items relevant to dark matter appeared during this period, including non-Newtonian gravity; X-ray emission from rich clusters and the X-ray background; a revival of gravitational lensing; discovery of the 3 K background and implications for galaxy formation from its extreme isotropy; two new dark matter candidates (neutrinos of non-zero rest mass and

primordial black holes); extended rotation curves; and Big Bang nucleosynthesis. Subsequent events that have changed the relationships of these items to the primary issue of existence and nature of dark matter are mentioned sporadically.

5.1 The 1961 Conferences

The International Astronomical Union General Assembly of 1961 was the first held in the USA and the first with associated symposia before and after. IAU Symposium 15 (McVittie 1961, McV below) ranged over the general territory of extragalactic astronomy and includes a brief remark by Zwicky. An additional associated meeting was specifically targeted at “The instability of clusters of galaxies” (Neyman et al. 1961, NPS below). Zwicky was not a speaker, was referenced in only one meeting presentation, and organized an unofficial evening session to present his own ideas, but then spoke largely in opposition to the idea of superclusters of galaxies rather than about dark matter, according to Vera Rubin, who was there.

A wide range of views was expressed at the two events, most of which had a few adherents up to the 1980s (Trimble 1995, [▶ Sect. 5](#)), but dark matter was, of course, the winner. Holmberg (in NPS) hoped that the problem would just go away and suggested that observational errors, foreground/background interlopers, and substructure in clusters could account for large velocity dispersions. Lemaitre (NPS, p. 603) proposed that rich clusters might continuously exchange galaxies with the field, so that the configurations were long-lived, but the individual members need not be gravitationally bound together.

Ambartsumian (NPS p. 536 and references therein) was firmly committed to the idea that clusters of galaxies were expanding out of some denser, more compact configuration, in somewhat the same way that some star clusters are clearly unbound and expanding. Supporters in 1961 included Vorontsov-Velyaminov (NPS p. 551), Kalloghjian (NPS p. 554), Markarian (NPS, p. 555), de Vaucouleurs (NPS p. 629), and the Burbidges (NPS p. 541).

Even then, the analogy with star clusters could have been only an approximate one, because other star clusters are clearly bound by the stars themselves, and very young ones are forming from gas clouds that were initially contracting but dissipate when kinetic energy is added by winds and supernovae of the most massive stars. And we see no $10^{17} M_{\odot}$ gas clouds that could be the immediate precursors of unbound clusters to form in the future. In fairness to Ambartsumian, he never supposed we did. His image was more nearly akin to that of Jeans, who once described the spiral nebulae as places where material was pouring into our universe from elsewhere. In fact, Ambartsumian’s image of stellar evolution was a similar one, where compact configurations came first, and dense pre-stellar matter expanded to make the stars (with planetary nebulae as an early phase rather than a late one in stars’ lives). A crucial issue is that rich cluster galaxies, especially in cores, are mostly ellipticals, and field galaxies more often of later types, a point emphasized by van den Bergh (1960) (NPS p. 566). He also briefly considered a possible role for non-gravitational forces (also the view of de Vaucouleurs 1960), but rejected it.

There was, however, also a “stability of clusters” coterie at the meetings. Zwicky (McV p. 347) and Baum (McV p. 255) thought there might be enough intracluster starlight to bind the clusters with luminous material. Minkowski (NPS p. 558) and Limber (NPS, p. 572) were also on the “bound side.” Most committed was Abell (NPS. P. 607). He had taken a large fraction of the Palomar Observatory Sky Survey plates, examined them in detail for clusters and other structures, and compiled a catalog of clusters (still in use). The process had persuaded him that not only clusters but “second-order clustering of galaxies and interactions between clusters of galaxies” existed, on length scales of 50 (100/H) Mpc and with masses of 10^{17} (100/H) M_{\odot} .

Zwicky was a firm opponent of superclustering, but some of his clusters (in an independent catalog also derived from POSS plates) had sizes of 40 (100/H) Mpc and enough substructure that an innocent bystander might well have said they displayed second-order clustering.

Van Albada (NPS p. 590) and von Hoerner (NPS p. 580), both better known for other things, attempted to determine the extent to which the Virial theorem could be applied to rich clusters when only a handful to a score of redshifts had been measured. Using analytical and numerical methods, respectively, they concluded that Virial masses were likely to be wrong, but not by factors of more than two or three after significant relaxation of the clusters. Von Hoerner's calculation was plausibly the first N-body simulation in astrophysics. It had $N = 16$.

5.2 X-ray Considerations

These give us numbers for both gas and total mass of rich clusters and limits on cosmic diffuse hot gas. The first extra-solar-system X-ray discoveries were Sco X-1 and a very roughly uniform background (Giacconi et al. 1962). The standard pioneering discussion of possible emission mechanisms is that of Felten and Morrison (1966), who considered among other possibilities thermal brehmsstrahlung of ionized hydrogen, in an amount that could, just about, close the universe (for H around 50) without violating the Gunn and Peterson (1965) limit on neutral hydrogen in ionization equilibrium with it. Other calculations established that, even if the entire background were due to hot, ionized hydrogen, the universe could not be closed that way (Field and Henry 1964; Gould and Burbidge 1963; Hoyle 1963). Ruling out a significant contribution was, however, a slow process, driven by the gradual increase in the fraction of the observed background attributable to resolved sources and the angular fluctuations of the rest (Gursky and Schwartz 1977, reporting when the issue was just about resolved).

In the rocket X-ray era, there were tentative detections of one or two clusters of galaxies as X-ray sources. The Uhuru satellite reported something like 40 in 1971–1974. Again the theorists considered all the usual radiation mechanisms, but in this case detection of an iron emission line settled the question in favor of thermal brehmsstrahlung. Early estimates of the amount of gas required in some cases came close to the total needed to bind the clusters, but it soon became clear that the gas mass is about 10% of the total, comparable with the mass in galaxies (again settled by the time of the Gursky and Schwartz 1977 review). But, of course, the hot gas must be contained through the life of the cluster, and the mass required to do that proved comparable with total cluster masses found from galaxy velocity dispersions and the Virial theorem, providing an independent demonstration of dark matter in rich clusters. An easy, approximate way to think of it is that kT/m_e is essentially the appropriate velocity for an $M = V^2 R/G$ mass, where R is some measure of cluster size. Whether the masses found in the two ways are the same, or should be, is mentioned in [Sect. 6](#) below.

5.3 Nonstandard Gravitations and Cosmologies

The possibility of non-gravitational forces had been mentioned briefly at the 1961 conference (cf. de Vaucouleurs 1960 and Hogan and White 1986 for a more recent version.) Non-Newtonian (or non-relativistic) gravity can be equivalent, and appears quantitatively first in the

work of Finzi (1963), who marshaled evidence for M/L values increasing outward in the Milky Way, and so suggested that the deviation was in the direction of G becoming stronger than $1/R^2$ at large distances. Currently viable variants of alternative gravity belong to a later period, while non-zero Λ never quite disappears from the astronomical recipe book (McCrea 1971).

In case any reader is shouting to himself “magnetic fields,” he might be reminded that these constitute a highly relativistic fluid. Thus if they are to confine anything, they must be firmly anchored to gravitating matter, which must be sufficient to contain the fields themselves as well as whatever they originally set out to hold on to.

5.4 Gravitational Lensing (and Deflection of Light)

The early literature of this topic is much richer than is generally recognized. Trimble (2001) cites 8 papers in German and 10 in English from before 1961 and 8 in Russian and at least 15 in English from 1961–1974. Among the items more or less understood were:

- There is no precise analog with bending (etc.) of light by material lenses, because it is not possible to have a region of constant index of gravitational refraction, but rays passing just on either side of the sun would meet outside the orbit of Neptune, making it an $F = 59,600$ system.
- The Newtonian/particle theory of light calculation was done by Soldner, Cavendish, and others before 1900, besides Michell and Laplace, and, of course, Einstein in 1907–1908.
- The extra factor of two bending in General Relativity constitutes a test of the theory.
- The possibility of both rings and multiple images in what are now called strong lensing was pointed out in the 1920s (by Oliver Lodge and O. Chowlson, of St. Petersburg but writing in German), and Edwin Frost at Mt. Wilson apparently looked for star-star lensing in 1923. (He didn’t find any.)
- Rudi W. Mandl drove both the interwar calculations by Einstein and by Zwicky.
- Zwicky (1937) understood that gravitational lensing could measure galaxy masses, magnify the images of distant sources, test general relativity, and was more likely for galaxies than for stars.
- H.N. Russell imagined what an inhabitant of a planet orbiting Sirius B might see, including lensed arcs from extended sources.
- Other well-known people dabbled in the territory (Charles Darwin, the grandson of the one you know, Ya. B. Zeldovich, James E. Gunn, P.J.E. Peebles, Philip Morrison and his students...)
- Weak lensing was also calculated with increasing precision, though generally viewed as a noise source – like “a shower door,” imposing fundamental limitations on the accuracy with which you can measure angular diameters of distant galaxies – and it was shown that the optical depth of the universe to weak lensing by any population is roughly the same as the fraction of closure density that population contributes.
- The effects of galaxies and clusters on the cosmic microwave background include lensing effects.
- Optical depths for star-star lensing were calculated by Lieves (1964) who found magnifications of 2,000 possible, but an optical depth in the Milky for even a factor 10^{-9} .

The first paper likely to be mentioned in recent discussions of gravitational lensing is, however, that of Refsdal (1964). This was a true “sleeping beauty” paper in the language of the community that analyzes citation rates. Almost ignored at the time of publication and for at least 15 years afterward, it is now widely known, perhaps most because the author pointed out you could use the multiple lensed images of variable QSOs or quasars (with known redshifts for both source and lens) to measure the Hubble constant. In the process, you also get a mass for the lens, as Zwicky (1937) promised.

Strong and weak lensing produce multiple images and distortions, respectively. The phrase microlensing can describe either variability in a strongly-lensed source due to individual stars or other lumps in the lensing galaxy or star-star lensing in and around the Milky Way. These things have now all been seen, starting only in 1979 (Walsh 1979), though Barnothy and Barnothy (1968) had suggested that all QSOs and quasars were lensed and intrinsically only as bright as Seyfert galaxies. The phenomena come into their own relative to dark matter in three ways: tests for the nature of the galactic halo DM from microlensing searches; detection of dark halo substructure in galaxies that lens, and measurements of cluster masses to compare with the numbers from X-ray data and the Virial theorem.

5.5 The Cosmic Microwave Background and Its Remarkable Isotropy

The discoverers of the CMB do not need any more citations (but for a whole book about the subject, see Peebles et al. 2009). This radiation is important for dark matter because it is so smooth across the sky that only non-baryonic DM can build the current inventory of galaxies in the time allowed.

Even the first papers noted that the temperature or intensity did not seem to vary on any angular scale, with an upper limit of 10%. That limit shrank monotonically with time, the dipole due to our motion popping out at a part in 10^3 in about 1970 (Conklin 1969; Henry 1971). Meanwhile, as it were, astronomers (e.g., Rees 1971) were trying to form galaxies and clusters from the gradual growth of density fluctuations established in the very early universe. Now the problem is as follows:

From recombination to the present is $z = 1,000$ to 0, meaning that linear perturbations can grow only about a factor of 10^3 . Thus $\Delta\rho/\rho = 1$ now requires $\Delta\rho/\rho 10^{-3}$ then. But, for adiabatic fluctuations $\Delta T/T = (1/3) \Delta\rho/\rho$, since photon number density scales as $T^{1/3}$. Thus, for galaxies and clusters formed entirely from baryons, it is predicted that the radiation, streaming freely after recombination should be lumpy on the sky at a level of 3×10^{-4} . It is not. Zeldovich (1972) considered the possibility of isothermal fluctuations, but the underlying physics to produce these was problematic.

Dark matter made of baryons does not help, but DM that interacts with photons only gravitationally could be a galaxy-saver, predicting CMB fluctuations of parts in 10^5 rather than parts in 10^4 . Density perturbations can grow before recombination to larger amplitudes and gas can stream into the previously assembled DM halos after recombination to become galaxies and all. The most prominent DM candidates belong to a later time frame, and the tip of the relevant CMB fluctuations were seen only with the launch of the COBE satellite (Smoot et al. 1992), with the ones directly connected to current structures belonging to the very recent epochs of WMAP and some preceding balloon-borne telescopes (Vol. 6)

5.6 The First Non-Baryonic Candidates

At least five of these appear in the literature before 1974. All are now thought to be likely, but not dominant, contributors. The existence of neutrinos was generally accepted from the mid 1950s and their masses understood to be zero or small. The laboratory limits were, however, not very tight, and Gershtein and Zeldovich (1966) realized that one could do better simply by insisting that the universe not be strongly over-closed by them. Their limit was about 400 eV for a single type of neutrino or antineutrino, and Ω less than 10. A more stringent limit on Ω and equal weight given to electron and mu neutrinos and their antiparticles led Cowsik and McClelland (1972) down to less than 20 eV, but within the range that could close the universe. Laboratory measurements of the mass differences among the three flavors now known are a good deal smaller and make them all probably less than 1% of Ω .

A second possibility, from the same Russian school in the same year (Zeldovich and Novikov 1966) was primordial black holes. This means ones formed before the epoch of nucleosynthesis, so that baryons could have gone into them without disturbing the production of helium and deuterium (► Sect. 5.9). Again the authors were not primarily thinking of PBHs as dark matter, and indeed they are not dark if there is gas around to be accreted and to radiate as it goes. In addition, BHs can bend light coming to us from sources behind them (gravitational microlensing for BHs of planetary and stellar masses), and for asteroidal masses around 10^{15} g perhaps boil away (Hawking radiation), while large ones (e.g., $10^6 M_{\odot}$) in galaxies will wreak havoc with star clusters, giant gas clouds, and even whole disks. Large numbers of $10^8 M_{\odot}$ or larger black holes would lens QSOs in a way not seen. These five processes enable the exclusion of most, but perhaps not quite all, masses of black holes as dominant dark matter candidates.

In a short, prescient paragraph, Sciama (1971, p. 129) mentions what he calls “missing material,” saying that the individual faint stars, rocks, neutrinos, or gravitational waves would not yet have been observed. Like PBHs, gravitational waves of most wavelengths would have revealed themselves in the interim (for instance by wiggling the clocks that are called pulsars), but there is possibly still some residual phase space. The longest wavelength gravitational waves are likely to be those left from the early universe (“primordial”), discussed by Grischuk (1974).

Yet another idea from fundamental physics is that of various space-time singularities, now called monopoles, strings, domain walls, and textures. These can be produced by spontaneous symmetry breaking during phase transitions (Kriznits and Linde 1972), and it was immediately clear that one does not want very many of any of them around (Zeldovich et al. 1974).

5.7 Rotation Curves Again

For whatever reason, this is the length scale on which dark matter first acquired mainstream acceptance, and even non-technical discussions of the topic typically begin by saying that the outer regions of galaxies would fly apart without the extra gravitational force from dark matter. The data, however, belong only just barely to the 1961–1974 period, and then only for M31, with improved optical data (Rubin and Ford 1970) and 21 cm measurements extending further out (Roberts and Rots 1973; Roberts and Whitehurst 1975). The key point is that the rotation speeds do not decline outside the optically bright portions of galaxies, implying that M/L is rising outward. Freeman (1970) was nearly alone in appreciating that the phenomenon might be widespread and would be important.

Again somewhat inexplicably, the optical rotation curves have been more widely recognized, so that, for instance, the conference proceedings edited by Kormendy and Knapp (1985) mentions Rubin on 37 pages and Roberts on only 4. Trimble (1987, [▶ Sect. 2.2](#)) is similarly guilty, though perhaps only at the 2:1 level.

5.8 Pure Theory: Disk Instabilities

A thin, self-gravitating disk is unstable to bar formation and various other distortions (sometimes seen in accretion disks as well as galaxies). Ostriker and Peebles (1973) are generally credited for pointing out the instability and for showing that a spheroidal halo of comparable mass could stabilize them. Infrared imaging has shown that there are more barred spirals than were thought at that time (Casasola et al. 2008).

5.9 Big Bang Nucleosynthesis

This topic is relevant because, by 1974, it could just about rule out a universe closed by baryons. George Gamow and his younger colleagues had tried to make the entire range of chemical elements from neutron matter in an early universe (heated by the decaying neutrons). This fails most conspicuously because there are no stable nuclides with $A = 5$ or 8 . Thus the next stage was “the early universe made hydrogen and helium, but Burbidge et al. (1957) made all the rest.” All four were, to varying degrees, steady state supporters, thus it is at least curious that the latter two participated in one of the early quantitative calculations of how much H and He you should get from a Big Bang (Wagoner et al. 1967). Wagoner (1973) improved the calculations, allowing for dependence on expansion timescale, neutron half-life and other uncertain nuclear physics, and baryon density, among other variables.

He/H is rather insensitive to most of those variables, so that the discovery of interstellar deuterium at a level $D/H = 1.5 \times 10^{-5}$ (Rogerson and York 1974) helped a good deal to tie things down. Thus Gott et al. (1974) could say that a cosmic (baryon) density in excess of about $\Omega = 0.15$ would yield more helium and less deuterium than is found locally. What you get depends separately on the baryon density (a matter of how easy it is for particles to find each other) and the total density (the expansion time available for them to do it) as pointed out, perhaps first, by Shvartsman (1969).

Gott et al. (1974) also invoked the M/L values of large clusters (perhaps 200 in solar units) and the ages of population II stars and radioactive elements vs. the reciprocal of the then-favored Hubble constant to conclude that the universe is open. These three items were all correct in the sense of revealing that the total density of matter in any form is considerably less than $\Omega = 1$. Though the paper was published on 1 April, Λ was not mentioned, but Gunn and Tinsley (1975) had a go at an $\Omega = 1$ universe with large values of both ρ and Λ which did not catch on at the time.

In the ensuing 35 years, holding by the requirement that the early universe must make the proper amounts of H, H^2 , He^3 , He^4 , and Li^7 has helped to rule out a wide range of non-standard cosmologies, including strong anisotropies and decaying particles, that belong to Vol. 6 of this series.

5.10 The 1974 Papers

Jaán Einasto (2001) has described in some detail the logical process that led up to the first of the two watershed papers. His own work on the largest available sample of binary galaxies was key, as was the insistence by Zeldovich that he and his colleagues submit to a major journal (Einasto et al. 1974). Their key points were “The mass of galactic coronae halos exceeds the mass of populations of known stars by one order of magnitude, as do the effective dimensions” and “The mass-luminosity ratio rises to $f \approx 120$ for elliptical galaxies. With $H = 50$ km/s/Mpc, this ratio for the Coma cluster is 170.” A careful reading of their paper suggests that they had meant to include a table or graph of M/L vs. length scale, which somehow got left out (Nature papers were still very short in those days).

Einasto et al. were just in time for the second paper (Ostriker et al. 1974) to cite their preprint. The latter concluded “Currently available observations strongly indicate that the mass of spiral galaxies increases almost linearly with radius to nearly 1 Mpc... and that the ratio of this mass to the light within the Holmberg radius, f , is $\sim 200 (M_{\odot}/L_{\odot})$. They show a graph of M/L v.s length scale which would extrapolate to the equivalent of closure density (M/L $\sim 1,000$, remember) at $R = 3,000$ Mpc, the Hubble radius. Over the years, the Ostriker paper has been cited about twice as often as the Einasto paper. It would be interesting to chase down a large sample of the citations and decide whether the main driver is simply American chauvinism.

Very soon after, Ozernoy (1974) pointed out that, within rich clusters, the dark matter must mostly belong communally to the cluster and not to the individual galaxies, first because the galaxy separations (at least in cluster cores) were smaller than the necessary DM halo sizes, and second because there is not as much luminosity segregation as would be expected from mass segregation if those hefty monsters were interacting gravitationally.

A first counterblast to this new dark matter paradigm came from Burbidge (1975) emphasizing, as Holmberg had done in 1961, selection effects, observational errors, and other uncertainties. He remained unreconciled to what is now conventional dark matter at least until 2001 (Narlikar 2001, Sect. 13).

6 Dark Matter and Standard Models Since 1974

Practicing scientists will normally put the cut between history and current events at the time when they started reading the literature for themselves, probably early in graduate school. This was about 1965 for the present writer, but could easily be 40 years later for some readers (and perhaps as much as 20 years earlier for a very few others). Some time in that 40 years, pervasive, at most weakly interacting, dark matter became part of customary astronomy and cosmology. This section is intended to bridge the gap from the previous ones to the considerations of modern cosmology that will appear in Volume 6. The main topics are the impact of inflation, the cosmological constant yet again, dark candidates, structure formation, residual doubts, and dark matter alternatives. These are not arranged chronologically, but perhaps in the order you might meet them in an introductory textbook.

6.1 The Impact of Inflation

The general idea is a very early (before 10^{-23} s or so) epoch of exponential cosmic expansion, followed by reheating to the high temperatures required for nucleosynthesis, thermalization of neutrinos, and so forth. Reserving the name Big Bang for that latter period would save a great deal of fuss and bother, because there is considerable observational evidence that it actually happened. Inflation solves several problems called causality, monopoles, flatness, large-scale homogeneity, and origin of small-scale inhomogeneities (▶ [Chap. 7](#) of Vol. 5) that had not bothered most astronomers until the particle physicists told us about them (compare telephone solicitations for aluminum siding and comprehensive telephone plans).

From the point of view of dark matter, the key property of inflation is that it guarantees that the universe will be very nearly flat ($k = 0$) and have the closure density in the sum of all forms of matter plus cosmological constant, because a very large increase in the scale parameter $R(t)$ in the relevant equation drives the k term to zero:

$$H^2 = \frac{8\pi G\rho_{\text{in}}}{3} - \frac{k}{R^2}$$

where H is the Hubble constant, G is the constant of gravity, ρ_{in} is the energy density of the scalar field invoked to drive the inflationary (exponential) expansion, and $c = 1$ in the units loved by cosmologists who do not build apparatus. Various forms of matter and radiation dominate Λ until rather recently, if Λ is truly a constant. The key papers normally cited are Guth (1981) and Linde (1982). A good didactic presentation is that of Kolb and Turner (1990, [Chapter 8](#)).

Thus the idea of inflation drove astronomers back to thinking about a critical-density universe, initially closed by some form of (dark) matter and later by some form of Λ , cosmological constant, dark energy, or quintessence.

6.2 The Final Resuscitation of Λ

The cosmological constant has hung around the fringes of the universe since 1917 Einstein (1917). He dropped it from his later cosmological papers, but the evidence that he claimed it as his greatest blunder is third-hand. Zeldovich (1968) pointed out that Λ could be thought of as a vacuum field energy density, for which the Lamb shift, Casimir effect, etc. provide independent evidence of a sort. But its re-incorporation into mainstream cosmology was driven by observations. Kofman et al. (1993) presented perhaps the first “ Λ CDM” universe, based on COBE measurements of the CMB anisotropy, large-scale clustering, and ages of things. Their Λ was about 0.8 of the critical density, with the rest in some combination of cold dark matter and baryons. By 1995, Peebles (1995) is ready to say that “the last column [of Table 1] assumes at the present epoch the mass density is subdominant to only one significant term, Λ or space curvature,” the latter meaning an open universe. And at an IAU Symposium (S183, Cosmological Parameters and Evolution of the Universe), held as part of the General Assembly in Kyoto in 1997, a panel discussion ended with a sort of vote in which the majority of the participants endorsed something like the present consensus model, with H around 70 km/s/Mpc, 5% of the closure density in baryons, another 20% or so in dark matter, and the rest in cosmological constant. Unfortunately, the discussion does not appear in the symposium proceedings, and three of the panel members who made up part of the majority were co-opted after the program book

was printed (J.P. Ostriker, the present author, and one other), so that the record is only in the memories of the participants and any surviving notes they might have taken at the time.

The yearly review, *Astrophysics in 1997* (Trimble and McFadden 1998) was the first in a series of 16 to record a majority of papers favoring non-zero Λ . Then, starting in 1998 came the supernova and WMAP data that completed the process of Λ incorporation (Spergel et al. 2007).

6.3 Candidates: Dark, Darker, and Darkest

Section 5.6 mentioned neutrinos, black holes, gravitational radiation, topological singularities (useful really only as seeds for galaxy formation); and the main baryonic candidates (very faint stars, stellar remnants, and gas in elusive density/temperature regimes) are left from Sect. 1. Many more recent ones have come from considerations in elementary particle physics. First of the two most generally invoked are weakly interacting super partners of known bosons, given names ending in “ino” (like photino and gravitino) or, more generally, WIMPs (Weakly Interacting Massive Particles, the name introduced deliberately by Steigman and Turner (1984)). The basic scheme was laid out by Lee and Weinberg (1977), showing that likely particle masses (GeV or more) and cross sections (10^{-42} cm² or less) indeed yield an appreciable, but nearly undetectable, cosmic density. The other entity, axions (Peccei and Quinn 1977), is co-eval, at least in the literature, and is also a possible form of cold dark matter, non-relativistic at the epoch of galaxy formation, because axions form out of a Bose–Einstein condensate, despite their masses being much less than an eV.

And then there are all the rest. The total number tabulated by Trimble (2005) is about 75, and the list is by no means exhaustive. Trimble et al. (2007) present 23, not all new. Some arise from front-line physics and could be part of a unified dark sector. Some favorites are SuperWIMPS (meaning even smaller cross sections), WIMPzillas with masses of 10^{15} GeV, and Kaluza-Klein particles. Some had brief periods of glory that quickly faded, like self-interacting dark matter (larger cross sections but only for particles encountering each other) and decaying dark matter with mass = 27 eV. And some are just odd, for instance DAEMONS (Dark Electric Matter Objects). These are particles for which thermodynamic time runs backward, preventing electromagnetic interactions.

The opposite of a WIMP is, of course, a MACHO, a MAssive Compact Halo Object in the Milky Way halo that could gravitationally lens stars in the Large Magellanic Cloud or in the disk that could lens stars in the bulge. The lensing is seen, and dividing up the lenses among known faint stars in the Milky Way and LMC and expected MACHOs (white dwarfs, neutron stars, black holes, other very faint entities) still presents some problems (Torres 2008). Whatever they are, however, they are not most of the mass in our halo. The original MACHO project completed its intended surveys (Alcock et al. 2003), though some related ones (called OGLE, ASAS, etc.) continue and are now finding large numbers of planets transiting their (non-lens) host stars and a few lenses with planets as well.

6.4 Galaxy and Cluster Formation with Dark Matter

No sooner had Lee and Weinberg (1977) codified the idea of supersymmetric partner particles than astrophysicists began incorporating them, as cold dark matter, into their models of

structure formation (Gunn 1977; White and Rees 1977). This was, from the beginning, biased cold dark matter, because the results of calculations resembled the real masses and spacings of galaxies only if the galaxies were restricted to the highest peaks of DM density, so that $\Delta L/L = b \Delta M/M$, with the bias factor $b = 1.5\text{--}2.5$. Blumenthal et al. (1984) is generally regarded as the definitive demonstration that biased CDM was beginning to solve “the galaxy formation problem,” at least on scales up to about 10 Mpc. And see Solovevo and Starobinskii (1985) for the view from further east.

On larger scales, however, the simulations did not produce nearly so many of the 30–100 Mpc voids, clusters and deviations from uniform Hubble flow as are seen in the real world. Let Bertschinger (1991) stand for very many papers reporting this problem, and Hamilton et al. (1991) for a comparable number pointing out that almost anything added – including baryons, hot dark matter, topological defects or other seeds for galaxy formation, a complex spectrum of initial perturbations, or, as is now recognized, nonzero Λ – would improve the situation.

Where was the hot dark matter while all this was going on? Lurking on the sidelines from the time when Tremaine and Gunn (1979) pointed out that the exclusion principle did not allow enough quantum states for neutrinos with realistic masses to bind the smallest dwarf galaxies with their shallow potentials. A few years later, HDM was to be found as an adjunct for producing the largest scale structures (Doroshkevich 1984; Umemura and Ikeuchi 1986). Recent work on galaxy formation is happy with Ω (HDM) = 0.006, about the number expected from the known neutrino mass differences (Allen et al. 2003).

Sterile neutrinos are ones that do not participate in either big bang nucleosynthesis or neutrino oscillations under current conditions. If they close the universe, they could be massive enough to be Warm Dark Matter (semi-relativistic) at the epoch of galaxy formation. Knebe et al. (2008) have reconsidered the possible role of WDM for galaxies.

Virtually all twenty-first century simulations of the assembly of galaxies and clusters, however, make use of a standard Λ CDM mix of ingredients. The Millenium Simulation (Springel et al. 2005) was the one to beat, and to use, a few years ago, and has indeed probably now been beaten by somebody, after a few additional “Moore’s Law quanta.” There are two issues: first the need to incorporate as many physical processes (including baryonic ones) as possible, and, second, the drive for ever-larger numbers of particles in the calculation, so as to be able to resolve structures of $10^6 M_\odot$ (dwarf galaxies; young globular clusters) or less. Incidentally, confidence in the main outlines has grown to the point where the presence or absence of dark matter is used to distinguish very small galaxies (with) from large star clusters (without, Fellhaver 2008; Mieske et al. 2008). The distinction between a large bathing machine and a very small second class carriage is left as an exercise for the reader.

6.5 Residual Doubts

Well into the “consensus era,” a small number of brave observers continued to report data sets that were, they said, more nearly, or at least as, consistent with $\Omega_M = 1$ or Ω (total) small (Trimble and Aschwanden 2003, Sect. 12.2). On the other hand, an earlier claim that dark matter should not be taken seriously, because rich clusters yielded three different numbers from the velocity dispersion, X-ray data, and weak lensing was resolved. Cohen and Kneib (2002), for instance looked at RX J1347-1145 in all three ways and concluded that the Virial mass was smallest, the X-ray mass largest, and lensing came in the middle. They suggested that the cause

might be a merger in progress, which has shocked the gas (making it hotter than it ought to be) while the two separate velocity dispersions have not yet had a chance to discover that they belong to a single, more massive cluster. Generally, non-thermal pressure support (shocks, magnetic fields, turbulence) will yield a too large X-ray mass; substructure and non-members can either enlarge or contract the velocity dispersion; and lensing is not the desired gold standard, because the distortions are subtle enough you have to postulate some spherical or other simple model for the lensing cluster to carry out the analysis.

At the other extreme from these almost mainstream reservations come the sorts of items you may well have received as part of a mass postal or e-mailing. The most recent on my desk comes from Tsiganov and Tsiganov (2009) and expresses the view that conventional gravitation physics went wrong somewhere around the time of Galileo and Huygens. The illusion called dark matter is, they say, one of very many consequences of this mistaken view of gravity. The authors are at the University of Economics and Law in Ukraine, but other papers almost as remarkable have come from departments of physics, astronomy, and engineering.

6.6 Dark Matter Alternatives

Now that Λ is part of the standard universe, this has come to mean theories of gravity other than general relativity and cosmologies other than Friedmann-Robertson-Walker-Lemaitre which wholly or partly replace dark matter. Many of these have short half-lives, are supported only by their originators, and have not been brought face-to-face with the full range of relevant data. This range includes solar system tests, the changing orbits of binary pulsars, the redshifts of X-ray lines from X-ray binaries with black hole accretors, and rotation of gas and velocity dispersions in galaxies, as well as structure formation, weak, strong, and microlensing, and the evolution of merging galaxy parts. Trimble (2005) summarizes half a dozen such theories. Veveschagin and Yegorian (2008) review a larger, partly overlapping set.

Closest to being fully worked out alternatives are MOND (MODified Newtonian Dynamics) and its relativistic extension (Bekenstein 2006; Milgrom 2007, and earlier papers referenced therein). The general idea is that of a minimum possible gravitational acceleration, below which gravity turns over to a $1/R$ rather than $1/R^2$ law. Of at least equal importance from the point of view of history of science is that a sizable number of astronomers, astrophysicists, etc. (at least a dozen) have taken the idea seriously enough to explore its consequences (S.M. McGaugh, 2009, personal communications kindly provided a list of the men (and a very few women) of MOND.) Thus, while a typical month of a major journal is like to have 5–10 or more papers exploring the properties and consequences of dark matter or assuming its existence for some other purposes like data analysis, there will also usually be one or two MOND papers. I happened to look at the June 2008 issues of MNRAS and the June 2009 issues of ApJ. Many of these current authors are younger than the originators of the idea, suggesting it will have a long half-life. In some cases, dark matter (conceivably sterile neutrinos) is needed even with MOND (Richter et al. 2008).

6.7 From the Standard Hot Big Bang to Consensus Cosmology

Both the Big Bang and inflation have passed a number of tests that they might have failed, coming from data on primordial abundances, CMB brightnesses and temperature isotropy on large

scales plus small fluctuations on small scales, measurements of ages of stars and radioactive elements vs. the Hubble constant, distant supernovae, and so forth. The present belongs to WMAP and its implications (Komatsu et al. 2009 on the fifth year data release).

Some unsolved problems, largely connected with small-scale structures, remain. Simulations tend to produce cuspy centers for galaxies and clusters, Vs. The nearly isothermal core seen (Siman and Geha 2007), and also more substructures in the halos of big galaxies than are seen as satellites (Abdelgudeu and Melia 2008; Madau 2008). Most researchers coming across these problems seem to think that better calculations, with more baryon physics and better mass resolution, are likely to solve them.

In summary, the question of the existence of dark matter has probably been subsumed by the consensus cosmology model. Its nature is a different story, given that there are very many candidates that cannot entirely be excluded and that the right one is perhaps still not in the inventory. How can progress be made? More and better calculations and observations of large scale structure and CMB distortions may be part of the story, but the happiest event would be some form of detection. There are three possibilities: observation of decay or annihilation products (photons, neutrinos, leptons) coming from the Galactic halo; capture in laboratory detectors; or production by particle accelerators. Several claims of the first two have appeared in recent years with at least partial refutations, addressed in Hooper and Boltz (2008) on DM and Caldwell and Kamionkowski (2009) on DE.

Acknowledgments

I am most grateful to Gerard Gilmore for the invitation to write this review and to Alison Lara for keyboarding it. The astronomers, some no longer with us, who have contributed to and influenced my thinking about dark matter range from Abell (Geroge Ogden) to Zwicky (Fritz), with special extra thank-yous to Vera Cooper Rubin and Daniel Magnus Popper for insights on “how things used to be.”

References

- Abdelgudeu, M., & Melia, F. 2008, MNRAS, 388, 1869
 Adams, J. C. 1847, Mem. RAS, 16, 427
 Alcock, C., et al. 2003, ApJ, 598, 597
 Allen, S. W., et al. 2003, MNRAS, 346, 593
 Babcock, H. W. 1939, Lick Observ Bull 19, 41
 Barnothy, J., & Barnothy, M. F. 1968, ApJ, 174, 477
 Bekenstein, J. D. 2006, Contemp Phys, 47, 387
 Bertschinger, E. 1991, in APS Conf. Proc. 272, After the First Three Minutes, ed. S. S. Holt et al. (New York: AIP), 297
 Bessel, F. W. 1845a, AN, 22, 145, 169 & 185
 Bessel, F. W. 1845b, MN, 6, 136
 Blumenthal, G., Faber, S., Primack, J., & Rees, M. 1984, Nature, 311, 587
 Bondi, H., & Gold, T. 1948, MN, 108, 252
 Burbidge, E. M., & Burbidge, G. R. 1961, AJ, 66, 541
 Burbidge, G. R. 1975, ApJ, 198, L7
 Burbidge, E. M., Burbidge, G. R., Fowler, W. A., & Hoyle, F. 1957, Rev Mod Phys, 21, 547
 Caldwell, R., & Kamionkowski, M. 2009, Annu Rev Nucl Part Sci, 59, 397
 Casasola, V., et al. 2008, A&A, 490, 61
 Cicaire, I., et al. 2008, AJ, 135, 2038
 Clarke, T. E., et al. 2003, ApJ, 601, 798
 Cohen, J. G., & Kneib, J. -P. 2002, ApJ, 572, 524
 Conklin, E. K., 1969, Nature, 222, 971
 Cowsik, R., & McClelland, J. 1972, PRL, 29, 669
 Da Rocah, C., & Mendes de Oliveira, G. 2005, MN, 364, 1069
 de Laplace, P. S. 1796, Exposition du System du Monde, Vol. II, (Paris) 305,
 de Vaucouleurs, G. 1958, ApJ, 128, 465
 de Vaucouleurs, G. 1960, ApJ, 131, 585
 Doroshkevich, A. G. 1984, Soviet Astron AJ, 28, 253

- Einasto, J. 2001, in ASP Conf. Ser. 252, Historical development of modern cosmology, ed. V. Martinez et al. (San Francisco, CA: ASP), 85
- Einasto, J., Kaasik, A., & Saar, E. 1974, *Nature*, 250, 309
- Einstein, E. 1917, *Sitz. könig. Preuss. Akad. Wissenschaften* (Berlin: Springer), 142–152
- Fasenko, R. 1985, *Astrofizika*, 20, 495
- Fellhaver, M. 2008, *MN*, 385, 1095
- Felten, J. R., & Morrison, P. 1966, *ApJ*, 146, 686
- Field, G., & Henry, R. 1964, *ApJ*, 140, 1002
- Finzi, A. 1963, *MN*, 127, 21
- Freeman, K. C. 1970, *ApJ*, 160, 811
- Gershtein, S. S., & Zeldovich, Y. B. 1966, *J Exp Theor Phys Lett*, 4, 175
- Ghez, A., et al. 2008, *ApJ*, 689, 1044
- Giacconi, R., et al. 1962, *PRL*, 9, 439
- Gott, J. R., Gunn, J. E., Schramm, D. N., & Tinsley, B. M. 1974, *ApJ*, 194, 543
- Gould, R., & Burbidge, G. R. 1963, *ApJ*, 138, 969
- Grischuk, L. P. 1974, *JETP*, 40, 409
- Gunn, J. E., & Peterson, B. A. 1965, *ApJ*, 142, 1637
- Gunn, J. E., & Tinsley, B. M. 1975, *Nature*, 257, 454
- Gunn, J. E. 1977, *ApJ*, 218, 592
- Gursky, E., & Schwartz, D. A. 1977, *ARA&A*, 15, 541
- Guth, A. R. 1981, *PRD*, 23, 347
- Halley, E. 1718, *Phil Trans Roy Soc*, 30, 736
- Hamilton, A. M. S., et al. 1991, *ApJ*, 374, L1
- Henry, P. S. 1971, *Nature*, 231, 516
- Hockey, T., et al. (Eds.) 2007, *The Biographical Encyclopedia of Astronomers* (New York: Springer)
- Hogan, C., & White, S. D. M. 1986, *Nature*, 321, 575
- Holmberg, E. 1937, *Lund Observ Annals*, 6, 173
- Holmberg, E. 1954, *Lund Medd Series I*, 186
- Hooper, D., & Boltz, F. A. 2008, *Annu Rev Nucl Part Phys*, 55, 293
- Hoskin, M. (Ed.) 1997, *Cambridge Illustrated History of Astronomy* (Cambridge/New York: Cambridge Univ. Press)
- Hoyle, F. 1948, *MNRAS*, 108, 372
- Hoyle, F. 1959, in *IAU Symp. 9, Paris Symposium on Radio Astronomy*, ed. R. N. Bracewel (Stanford: Stanford Univ. Press)
- Hoyle, F. 1963, *ApJ*, 137, 993
- Hubble, E. P. 1934, *ApJ*, 76, 44
- Jeans, J. H. 1922, *MNRAS*, 82, 122
- Kahn, P., & Woltjer, L. 1959, *ApJ*, 130, 105
- Kapteyn, J. C. 1922, *ApJ*, 55, 302
- Knebe, A., et al. 2008, *MN*, 386, 1029
- Kofman, L. A., Gnedin, N. Y., & Bahcall, N. A. 1993, *ApJ*, 413, 1
- Kolb, E. W., & Turner, M. S. 1990, *The Early Universe* (Reading: Addison Wesley)
- Komatsu, E., et al. 2009, *ApJ Suppl*, 180, 330
- Kormendy, J., & Knapp, J. (Eds.) 1985, in *IAU Symp. 117, Dark Matter in the Universe* (Dordrecht: Reidel)
- Kragh, H. 1996, *Cosmology and Controversy* (Princeton: Princeton Univ. Press)
- Kragh, H. S. 2007, *Conceptions of the Cosmos* (Oxford/New York: Oxford Univ. Press)
- Kriznits, D. A., & Linde, A. 1972, *Phys Lett*, 42B, 471
- Kuhn, T. 1962, *The Structure of Scientific Revolution* (Chicago, IL: Univ. Chicago Press)
- Kunth, R. 1952, *Zeitschrift fuer Astrophysik*, 28, 234
- Kuzmin, G. G. 1952, *Tartu Astron Obser Publ*, 32, 211
- Lalleman, A. 1960, *PASP*, 72, 76
- Lee, B., & Weinberg, S. 1977, *PRL*, 38, 2237, *PRL*, 39, 165
- Leverrier, U. J. J. 1847, *AN*, 25, 53
- Leverrier, U. J. J. 1848, *AN*, 580
- Lieves, S. 1964, *PR*, 133, B858
- Lindblad, B. 1926, *Uppsala Medd*, 3, 1637
- Linde, A. D. 1982, *Phys Lett*, 108B, 389
- Lohman, W. 1956, *Zeitschrift fuer Physik*, 144, 66
- Madau, P., et al. 2008, *ApJ*, 679, 1261
- Mayall, N. U., & Aller, L. H. 1960, *PASP*, 52, 278
- McCrea, W. H. 1971, *Quarterly J Royal Astron Soc*, 12, 40
- McVittie, G. C. (Ed.) 1961, *IAU Symp. 15, Problems in Extragalactic Research* (New York: Macmillan)
- Michell, J. 1768, *Phil Trans RS*, 6, 243
- Michell, J. 1784, *Phil Trans RS*, 74, 35
- Mieske, S., et al. 2008, *A&A*, 487, 921
- Milgrom, M. 2007, *ApJ*, 667, L45
- Narlikar, J. 2001, in ASP Conf. Ser. 252, Historical Development of Modern Cosmology, ed. V. Martinez et al. (San Francisco, CA: ASP), 175
- Neyman, J., Page, T., & Scott, E. (Eds.) 1961, *AJ*, 66, 537 (Conf. on Instability of Clusters of Galaxies)
- Newcomb, S. 1906, *Sidelights on Astronomy: Essays and Addresses* (New York: Harper & Brothers)
- Oort, J. H. 1927, *BAN*, 3, 275
- Oort, J. H. 1932, *BAN*, 6, 249
- Oort, J. H. 1940, *ApJ*, 91, 227
- Oort, J. H. 1965, in *Galactic Structure* (Vol V of the Kuiper compendium), eds. A. Blaauw & M. Schmidt (Chicago, IL: Univ. Chicago Press)
- Opik, E. 1915, *Bull de la Soc Astr de Russie*, 21, 150
- Opik, E. 1922, *ApJ*, 55, 406
- Ostriker, J. P., & Peebles, P. J. E. 1973, *ApJ*, 186, 467
- Ostriker, J. P., Peebles P. J. E., & Yahil, A. 1974, *ApJ*, 193, L1
- Ozernoy, L. M. 1974, *A Zh* 51, 1108 (Soviet Astron *AJ*), 18, 654)
- Page, T. L. 1952, *ApJ*, 116, 23
- Page, T. L. 1956, *ApJ*, 136, 685

- Page, T. L. 1960, *ApJ*, 132, 910
- Peccei, R. D., & Quinn, H. R. 1977, *PRL*, 38, 1440
- Peebles, P. J. E. 1995, in *ASP Conf Ser.* 88, *Clusters, Lensing, and the Future of the Universe*, eds. V. Trimble, & A. Reisnegger (San Francisco, CA: ASP), 1
- Peebles, P. J. E., et al. (Eds.) 2009, *Finding the Big Bang* (Princeton: Princeton Univ. Press)
- Pigott, E. 1805, *Philos Trans R Soc* 95, 131
- Pooley, D., et al. 2008, *ApJ*, 697, 1892
- Read, J. I., et al. 2006, *MNRAS*, 371, 885
- Read, J. I., et al. 2008, *MNRAS*, 389, 104
- Rees, M. J. 1971, in *Italian Physical Society, Proceedings of International School of Physics, "Enrico Fermi"*, Course 47, *General Relativity and Cosmology*, ed. B. K. Sachs (New York: Academic Press)
- Refsdal, S. 1964, *MNRAS*, 128, 295, & 307
- Richter, P., et al. 2006, *A&A*, 445, 827
- Richtrler, T., et al. 2008, *A&A*, 478, L23
- Roberts, M. S., & Rots, A. H. 1973, *A&A*, 26, 483
- Roberts, M. S., & Whitehurst, R. N. 1975, *ApJ*, 201, 377
- Rogerson, J. G., & York, D. G. 1974, *ApJ*, 186, L95
- Rubin, V. C., & Ford, W. K. 1970, *ApJ*, 159, 379
- Russell, H. N., Dugan, R. S., & Stewart, J. Q. 1926, *Astronomy* (Boston: Ginn & Co.)
- Schaeberle, J. M. 1896, *AJ*, 17, 37
- Schmidt, M. 1956, *BAN*, 13, 14
- Schwarzschild, M. 1954, *AJ*, 59, 273
- Sciama, D. 1971, *Modern Cosmology* (Cambridge: Cambridge Univ. Press)
- Shvartsman, V. F. 1969, *JETP*, 9, 184
- Siman, J. D., & Geha, M. 2007, *ApJ*, 670, 313
- Slipher, V. M. 1914, *Lowell Observ Bull.*, 2, 65
- Smith, S. 1936, *ApJ*, 83, 23
- Smoot, G., et al. 1992, *ApJ*, 396, L1
- Solovevo, L. V., & Starobinskii, A. A. 1985, *Sov Astron AJ*, 29, 367
- Spergel, D. N., et al. 2007, *ApJ Suppl*, 170, 377
- Springel, V., et al. 2005, *Nature*, 435, 629
- Steigman, G., & Turner, M. S. 1984, *Nucl Phys*, B252, 73
- Tinsley, B. M., & Larson, R. B. (Eds.) 1977, *The Evolution of Galaxies and Stellar Populations* (New Haven: Yale University Observatory)
- Tolman, R. C. 1934, *Relativity, Thermodynamics, and Cosmology* (Oxford: Oxford Univ. Press)
- Torres, S. 2008, *A&A*, 486, 427
- Tremaine, S. 1987, in *Dark Matter in the Universe*, ed. J. Kormendy, & G. R. Knapp (Dordrecht: Reidel), 547
- Tremaine, S. D., & Gunn, J. E. 1979, *PRL*, 42, 407
- Trimble, V. 1987, *Annu Rev Astron Astrophys*, 25, 425
- Trimble, V. 1988, in *Modern Cosmology in Retrospect*, ed. R. Bertotti et al. (Cambridge/New York: Cambridge Univ. Press), 355
- Trimble, V. 1995, in *Amer. Inst. Phys. Conf. Proc.* 336, *Cosmic Abundances*, eds. S. S. Holt, & C. L. Bennett (New York: AIP), 57
- Trimble, V. 1996, *PASP*, 108, 1073
- Trimble, V. 2001, in *ASP Conf. Ser.* 237, *Gravitational Lenses*, eds. T. G. Brainerd, & C. S. Kochanek (San Francisco, CA: ASP), 1
- Trimble, V. 2005, in *Neutrinos and Explosive Events in the Universe*, eds. M. M. Shapiro et al. (Dordrecht: Springer), 181
- Trimble, V., & Aschwanden, M. A. 2003, *PASP*, 115, 514
- Trimble, V., & McFadden, L. 1998, *PASP*, 110, 223
- Trimble, V., Aschwanden M. A., & Hansen, C. J. 2007, *Space Sci Rev*, 132, 1
- Tsiganov, E. P., & Tsiganov, O. E. 2009, *Nauka I Studia*, Vol. 2
- Umemura, M., & Ikeuchi, S. 1986, *Astrophys Space Sci*, 119, 243
- Valentijn, E. A. 1990, *Nature*, 346, 152
- van de Hulst, H. C., et al. 1957, *BAN*, 14, 1
- van den Bergh, S. 1960, *MNRAS*, 121, 387 and *ApJ*, 131, 558
- van den Bergh, S. 2001, in *ASP Conf. Ser.* 252, *Historical Development of Modern Cosmology*, eds. V. Martinez et al. (San Francisco, CA: ASP), 75
- Veveschagin, G. V., & Yegorian, G. 2008, *Int J Mod Phys*, D17, 203
- Wagoner, R. V. 1973, *ApJ*, 179, 343
- Wagoner, R. V., Fowler, W. A., & Hoyle, F. 1967, *ApJ*, 148, 3
- Walsh, D. R., Carswell, R., & Weymann, R. 1979, *Nature*, 279, 38
- Wheeler, J. A. 1968, *Am Sci*, 56, 1
- White, S. D. M., & Rees, M. J. 1977, *MN*, 183, 341
- Zabrenowski, M. 1986, *Astrophys Space Sci*, 117, 179
- Zeldovich, Y. B. 1968, *Sov Phys Uspekhi*, 11, 381
- Zeldovich, Y. B. 1972, *MNRAS*, 160, 1
- Zeldovich, Y. B., & Novikov, I. D. 1966, *A Zh* 43, 750, *Sov Astr AJ*, 10, 602
- Zeldovich, Y. B., et al. 1974, *JETP*, 40, 1
- Zwicky, F. 1933, *Helvetica Phys Acta*, 6, 110
- Zwicky, F. 1937, *ApJ*, 86, 217
- Zwicky, F. 1951, *PASP*, 63, 61
- Zwicky, F. 1958, *L'Astronomic* 72, 285

Index

A

Aberration of starlight, 1095
Absolute brightness, 244
Absorption, 590, 591, 596, 597, 600–610, 612, 615–618, 622, 628, 635, 1098, 1099, 1102, 1103
Absorption lines, 590, 591, 597, 600, 602–604, 606–610, 612, 617, 618, 622, 628
Abundance
 α -elements, 82, 84, 85, 95, 106
 gradients, 376–379, 386
 helium, 79, 94
 iron-peak elements, 84–87
 lithium, 58, 59, 79–81, 94, 101
 modification, 71–72
 neutron-capture elements, 58, 87–94, 101
 patterns, 58, 59, 61, 72–78, 90, 92–94, 101–103, 108
Acceleration, 730, 731, 734–736, 738, 751, 769, 772, 773, 781
Accreted satellite, 9, 13
Accretion, 500, 515, 523, 535–537, 541, 546
Active galactic nuclei (AGN), 246, 252
Advanced CCD Imaging Spectrometer (ACIS), 251
AGB stars. *See* Asymptotic giant branch (AGB) stars
Age-metallicity relationship, 382–385
Ages, 1104, 1110, 1112, 1116
AGILE, 791, 796, 823
AGN. *See* Active galactic nuclei (AGN)
All-particle energy spectrum, 730, 776
All-sky survey, 592, 595, 600
 Al_2O_3 , 530, 537
Alpha Centauri, 1096
Alpha elements (α element), 6, 77, 82, 84, 85, 92, 95, 106
 alpha-capture, 24, 32, 39
Amorphous carbon, 503, 504, 510, 513, 515, 528
Andromeda galaxy, 682
 radio emission, 644
Andromeda nebula, 1097
Antideuterons, 747–750
Antiprotons, 747–750, 783
Arecibo, 593, 605, 626
Argelander, F.W.A., 1096
Ar XVII, 253
ASCA, 251, 252
ASTRO-H, 825
Astrophysical journal, 1100

Asymptotic giant branch (AGB) stars, 514, 515, 528–531, 543, 545
Australia Telescope Compact Array (ATCA), 605, 612
Axions, 1113

B

B68, 502
“Baade’s Window,” 10, 11, 13
Baade-Wesselink method, 833
Balance, 572, 584
Balloons, 729, 730, 732, 734, 741–743, 765, 783, 784
Bar, 925, 937–941, 943–946, 948, 957–968, 974, 976
Bar formation, 1110
Barred galaxies
 magnetic fields, 675, 680, 690–693
 radio emission, 680, 691
Barred spirals, 1110
Baryon, 1094, 1098, 1103, 1108–1110, 1112, 1114, 1116
BDs. *See* Brown dwarfs (BDs)
Bending wave, 950–953, 975, 976
Beppo-SAX, 251
Bessel, F.W., 1093, 1095, 1096
Betatron acceleration, 532–534
BG. *See* Big grains (BG)
BHB. *See* Blue horizontal branch star (BHB)
Bias correction, 837, 843, 850, 851, 854, 865–867, 869
Big bang, 1112, 1115–1116
Big bang nucleosynthesis, 1105, 1110, 1114
Big grains (BG), 505, 511, 523, 532
Binary galaxies, 1102, 1103, 1111
Binary population, 142, 143, 179, 184, 233
Binary pulsars, 1115
Binary stars, 1062–1064, 1095, 1110
Blaauw, A., 244
Black holes, 252, 253, 261–267, 1093, 1094, 1105, 1109, 1113, 1115
BL Lac, 256
Blue horizontal branch star (BHB), 603
Bohm diffusion, 799, 802, 822
Bose-Einstein condensate, 1113
Bremsstrahlung, 806, 811, 815–817
Brown dwarfs (BDs), 118, 120, 127, 136, 143, 160, 161, 177–186, 190, 193, 198, 234
Burbidge, Margaret and Geoffrey, 1102

- Bulge(s), 7, 9–14, 16, 988, 989, 995, 996, 998, 1002, 1003, 1006–1017, 1023, 1024, 1026, 1027, 1029, 1030, 1033, 1034
- Burbidge, Geoffrey, 1102
- Burbidge, Margaret, 1102
- C**
- Ca II, 603, 604, 606–609, 612
- Ca IIK, 604, 612
- Calorimeters, 742, 743, 759
- CANGAROO, 795, 796, 799
- Canum Venaticorum, 1104
- Carbides, 515, 527, 537
- Carbon abundances, 30
- Carbon dust, 507, 515, 528, 531
- Carbon-enhanced metal-poor stars (CEMP stars), 60, 66, 67, 84, 96, 102, 103
- Carbon-rich stars, 60, 90, 99, 103, 104
- Cas A, 514, 528, 796, 804, 805, 808, 814, 815, 817, 818
- Casimir effect, 1112
- Catalogue, 590, 592, 595, 597, 598, 603
- Causality, 1112
- Ca XIX, 253
- CD-38°, 245, 63, 66, 67
- CDM. *See* Cold dark matter (CDM)
- CEMP stars. *See* Carbon-enhanced metal-poor stars (CEMP stars)
- Cepheids, 833–854, 856–862, 864
 - classical, 835–840, 857–860, 862
 - type II, 833, 835, 839–841, 856, 860, 861, 864
 - variables, 244
- Ceres, 1094
- Chandra* survey, 251–255, 260, 263, 793, 799, 800, 805, 815, 821
- Charge, 501, 523–527, 532, 538, 540
- Charge distribution function, 523, 526–527
- Chemical abundances, 23–52, 59, 61, 62, 64, 69–79, 95, 101
- Chemical enrichment, 57–110
- Chemical evolution, 57, 59, 75, 79–96, 102, 108, 110
- Cherenkov, 734, 743, 752, 753, 755–761, 763, 766, 770, 773
- Cherenkov telescope array (CTA), 825
- Chile, 261
- Chinese LAMOST, 109
- Chronometer, 98–100
- CHVCs. *See* Compact HVCs (CHVCs)
- Chwolson, O., 1095
- Circumnuclear rings, 249, 260
 - magnetic fields, 691
- Cirrus, 512, 523
- Classical bulges, 10, 13
- Classical cepheids, 835–840, 857–859, 862
- “Classical” dwarf galaxies, 77, 95
- Closure density, 1104, 1107, 1111, 1112
- Cloud random velocities, 552
- Clouds, 500–504, 513, 525–527, 535, 537, 539–542, 546
- Cluster ages, 96, 105
- Cluster dynamical evolution, 357–360
- Cluster mass functions, 357, 361–362
- Cluster of galaxies, 1105
- Cluster surveys, 350, 351
- CMD. *See* Color-magnitude diagrams (CMD)
- CO, 616
- COBE, 1108, 1112
- Cold dark matter (CDM), 100, 105, 106, 1073, 1076, 1077
- CO-line emission, 990, 996, 1032
- Collisional rates, 524–525
- Collisionless shock waves, 805
- Color-magnitude diagrams (CMD), 350–356, 360, 363, 374, 376
- Column density, 550, 552, 556–559, 563, 565, 566, 579, 583
- Column density distribution, 596
- Coma, 1099, 1100
- Coma cluster, 1100, 1104, 1111
- Common stars, 850, 852–853
- Compact HVCs (CHVCs), 602
- Complex A, 600, 610, 611, 616, 618, 632, 635
- Complex C, 610–612, 615, 618, 632, 635
- Complexes, 590–600, 602, 609, 610, 613, 618, 623, 627, 631, 635
- Complex GP, 595, 600, 611, 631
- Composition, 274, 283, 284, 293, 295–306, 315, 317–319, 328, 330, 334, 335, 337, 340, 501, 505–507, 513–516, 528–532, 535, 541, 542
- Composition in knee region, 769–773
- “Concordance” cosmological model, 15
- Condensation, 529–531, 535–537
- Convection, 729, 734, 735, 739, 741, 744–746
- Cooling rate, 578, 579
- Core collapse, 882–884, 892, 902–903, 908–913, 915
- Core-collapse SNe, 84, 88, 90, 92, 93, 96, 101, 102
- Cores
 - HVC, 501, 507, 537
- Coronal gas, 618, 628, 631, 633–635
- Corundum, 530, 537
- Cosmic distance scale, 244
- Cosmic microwave background, 1107, 1108
- Cosmic rays, 501, 532, 533, 541, 542, 545, 546, 643–649, 651, 654, 657, 658, 670, 674–676, 678–680, 696, 697, 700, 703–707, 709, 710, 712, 727–784
 - anisotropy, 731, 740, 746, 770, 776, 779–783
 - ankle, 728, 731, 773, 774, 776, 778, 781
 - injection, 797, 808, 809, 821, 823

- knee, 728, 730, 731, 734, 736, 742, 762–773, 776, 781, 784, 794, 795, 819, 821
- Cosmo-chronometry, 96–100
- Cosmogenic neutrinos, 778
- Cosmogony, 61, 100–108
- Cosmological constant, 1104, 1111, 1112
- Cosmological lambda-cold dark matter (Λ -CDM) simulation, 2, 16, 106
- Cosmological simulations, 60, 100, 110
- Cosmology, 1075–1081
- Coulomb collisions, 805
- Counter-earth, 1093, 1094
- Critical metallicity, 75, 103, 104
- CS 22892-052, 98, 99
- CTA. *See* Cherenkov telescope array (CTA)
- Cusp, 1068, 1069, 1071–1073, 1076, 1077
- Cyclotron, 265
- D**
- Dark energy, 1095, 1112
- Dark globule, 502
- Dark halo, 988, 989, 995, 1002, 1006, 1007, 1010–1013, 1017–1019, 1021–1030, 1032, 1034, 1035
- Dark matter, 735, 742, 747–749, 783, 988, 989, 1003, 1017, 1021, 1022, 1024, 1026–1029, 1034, 1035, 1041–1081, 1093–1116
core, 1044, 1062, 1065, 1069, 1071–1074, 1076–1078
- Dark matter halos, 602, 633, 634
- Dark stars, 1093, 1097, 1098
- Darwin, C., 1107
- David–Greenstein mechanism, 643
- Degenerate matter, 1099
- Delta Cephei, 1094
- δ Sct stars, 847, 862, 864
- δ Sct variables, 847
- δ SX phe variables, 847
- Density fluctuations, 1108
- Density of states, 518, 519
- Depletions, 500, 501, 506–507, 535–538
- Destruction, 501, 532–537, 545
- Deuterium, 610, 1109, 1110
- Deviation velocity, 590, 591, 593, 594, 597, 613, 614
- Diffuse
interstellar medium (ISM), 245, 248, 250–260, 262, 263, 267, 268, 644, 647, 650, 652, 653, 657, 675, 679, 704, 712, 790, 791, 823
- Diffuse sources, 250
- Diffusion coefficient, 730, 736, 739–741, 744, 746, 783
- Diffusive models, 735, 744, 745, 783
- Diffusive shock acceleration (DSA), 795, 799, 806, 815
- Dip model, 731, 776, 778
- Dipole, 1108
- Dirac, 1104
- Disc scale length, 858
- Disc shocking, 915–916, 918
- Disk, 244, 265–268, 514, 518, 525, 527–530, 546
- Disks and bulges, 12
- Dispersion relation, 931–935, 947, 951, 952
- Distance scales, 831–871
- Distribution function, 519–523, 526–527
- Domain walls, 1109
- Doppler boosting, 267
- Drag, 532, 533
- DSA. *See* Diffusive shock acceleration (DSA)
- dSph. *See* Dwarf Spheroidal Galaxies (dSph)
- Dust, 500–508, 510, 512–517, 522–524, 527–546, 605–609, 615–617, 629, 630
- Dust budget, 528, 529, 531
- Dust compounds, 507, 514, 516
- Dust cooling, 104
- Dust formation, 516, 528–531
- Dust grains, 643–646, 657, 672, 673, 708
- Dust mass, 507, 513, 528, 546
- Dwarf cepheids, 847
- Dwarf galaxies, 33, 46, 48–52, 61, 75–77, 94–97, 104–106, 108–110, 1100, 1114
- Dwarf spheroidal galaxies, 706–708
radio emission, 706, 708
- Dwarf spheroidal galaxies (dSph), 75, 77, 95
- Dwingeloo telescope, 592, 595
- Dynamical center, 248, 250, 262
- Dynamical friction, 262, 955, 966–969, 973, 974
- Dynamical models, 1057, 1062, 1065, 1067, 1071, 1072
- Dynamics, 987, 1021, 1029, 1034
- Dynamo action, 654, 656, 657, 680, 690, 692, 694, 700, 710
- E**
- Early-type stars, 262, 263
- Early universe, 58, 61, 66, 76, 84, 87, 89, 92, 100–104, 108, 110
- EBHIS survey, 593
- Eclipsers, 850, 860, 861
- Eddington, A.S., 1096
- Eddington limit, 260
- Edge-on galaxies, 673, 678, 685, 695, 696, 698, 699
- Effective temperature, 25, 26, 28, 39, 51
- Effelsberg, 593, 605, 626
- Einstein, E., 1095, 1104, 1107, 1112
- Electron recombination, 525, 526, 538
- Electrostatic grain potential, 527
- Elemental abundances, 506, 507, 512–514, 530, 535, 729
- Elemental depletions, 507, 535, 536

Elliptical galaxies, 680, 706–708, 1103, 1111
 radio emission, 706, 708

Elongation rate, 753, 754

ELS, 4–6

Emission intensity, 521

Emission lines, 249, 262, 263

Energy distribution function, 520, 521

Equivalent width, 604, 606, 607, 612

European extremely large telescope, 109

European Southern Observatory, 261

Event horizon, 258, 260, 265, 266

Exponential cosmic expansion, 1112

Extensive air shower, 729, 730, 733, 734, 752–758

Extinction, 500–504, 507–513, 523

Extinction cross section, 507–510, 513

Extinction curve, 502, 503, 509–511, 513, 523

Extinction efficiency, 507

Extragalactic origin, 736, 776

F

Faraday depolarization, 648–653, 661, 664, 678, 683

Faraday rotation, 644, 645, 649–653, 662, 664–668,
 671, 678, 680, 681, 684, 686–690, 694,
 699, 705, 707, 708, 710, 711

Far-ultraviolet rise, 523

Far-ultraviolet spectroscopic explorer (FUSE), 596,
 609, 616, 617

Far-UV, 502, 503, 509, 511, 513

Far-UV extinction, 509, 511

Fe $K\alpha$ line, 252, 253

Fe-line, 251, 254

Fermi acceleration, 501, 532, 740, 776

Fermi LAT, 792, 796, 806, 816, 818

Fe XXV $K\alpha$ line, 251

Field halo stars, 73, 94, 96, 100, 106

Field of streams, 13, 14

Field stars, 61, 63, 73–75, 81, 94, 96, 98–100

Filaments, 245, 247, 248

Flare(s), 258, 265–267

Flatness, 1112

Flocculent galaxies, 680, 693–695
 radio emission, 694

Fluorescence, 734, 752, 753, 757, 760–763, 770,
 773–775, 784

Fokker–Planck method, 892

Fornax dwarf spheroidal, 847, 849, 859, 862–864

Fragmentation, 632–633

Frame dragging, 267

Free-free emission, 249

Freeze out, 531, 541

Friedmann–Robertson–Walker–Lemaitre, 1115

Fullerenes, 544, 545

FUSE. *See* Far-ultraviolet spectroscopic
 explorer (FUSE)

G

Gaia, 109

Galactic bulge, 24, 41–43, 51, 250, 263, 273–341

Galactic center, 659, 664–667, 671, 672, 705
 radio emission, 664

Galactic center distance, 834–835, 854–858

Galactic chemical enrichment (GCE) models, 74

Galactic coordinates, 244, 247, 252, 254

Galactic disk, 349, 352, 367, 370–372, 374–377, 379,
 382, 386–388, 988, 990, 993, 997, 1000,
 1004, 1009, 1010, 1017, 1019, 1021,
 1025, 1098

Galactic fountain, 589, 628–630, 635

Galactic halo, 35, 36, 43–50, 52, 1010

Galactic nucleus, 244–269

Galactic plane, 245–248, 251, 253, 255, 257, 1097

Galactic plane surveys, 555, 576

Galactic rotation, 834–835, 847, 857, 868, 1097, 1099

Galactic structure, 349–352, 372, 376

 central hole, 466

 galactocentric radius R_0 , 458, 468

 inner bar, 465–466

 long bar, 450, 457, 459, 464–465, 473

 spiral structure, 460–462, 467

 stellar scale length, 459–460

 stellar warp, flare, cutoff, 462–463

Galactic thick disk, 8, 9, 16

Galactic warp, 575

Galaxy(ies), 727, 729–731, 734, 735, 738, 746, 751,
 772, 781, 783, 986–1035

Galaxy halos, 652, 655, 708

 radio emission, 652

Galileo, 1115

Galloway, T., 1096

Gamma ray astronomy, 790–794

Gamma-ray lines, 790, 791

Gamma-rays morphology, 812

Gamma-ray space telescope, 791

Gas, 551–553, 558–565, 567, 569, 570, 572–575,
 577–580, 582, 584

Gas and dust budgets, 528

Gas drag, 532

Gas models, 892, 894

GASS survey, 592

GCs. *See* Globular clusters (GCs)

G-dwarf problem, 631

General relativity, 1095, 1107, 1115

Generating function, 133, 135–137, 198

GHRS. *See* Goddard High Resolution Spectrograph
 (GHRS)

Giant Magellan Telescope, 109

Globular clusters (GCs), 2, 3, 5, 7, 9, 13, 15–16, 23,
 24, 29, 32, 33, 35, 43, 46, 48–49, 52, 60,
 61, 72–74, 81, 94–96, 105, 106, 118, 124,
 164, 166, 171–173, 184, 196, 210, 211,
 215–219, 232, 244, 839, 841–844, 846,

- 847, 849, 854, 863, 1096, 1100, 1102, 1103, 1114
- Goddard High Resolution Spectrograph (GHRS), 617
- Goodricke, J., 1093
- Gould Belt, 824
- Grain charge distribution, 526
- Grain chemistry, 542
- Grain-grain collision, 534
- Grain potential, 527, 532
- Grain size distribution, 500, 507, 509–512, 517, 523, 533, 538, 541
- Grain surface chemistry, 501, 542
- Graphite, 509–511, 513–515, 517, 521, 522, 527, 533, 534
- Graphite grains, 509, 517
- Gravitation, 988, 1002–1004, 1010, 1011, 1026, 1028, 1029, 1033, 1035
- Gravitational lensing, 1095, 1101, 1104, 1107–1108
- Gravitational microlensing, 1109
- Gravitational redshift, 267, 1095, 1099
- Gravitational waves, 1109
- Gravothermal catastrophe, 901–902
- Gravothermal oscillations, 903–904
- Gregory, J., 1096
- Group velocity, 934–935, 938, 943, 952, 953, 956
- Gunn, J.E., 1097, 1107
- GZK effect, 731, 732
- H**
- H α , 600–601, 614, 615, 623, 626
- HAC. *See* Hydrogenated amorphous carbon (HAC)
- Hadronic, 792, 797–799, 802–814, 819, 821–823, 825
- Hadronic interaction models, 754, 764, 769, 770, 775, 778
- Hadronic models, 798, 802–807, 814, 819, 822, 825
- Hadronic shower, 754, 760
- H α emission, 996
- Halo, 5–9, 12–17, 60, 61, 63, 66, 68, 72–76, 78, 94–97, 100, 101, 103, 105, 106, 108–110
- Hamburg/ESO survey (HES), 63, 64, 67, 73, 98
- HAWK, 793
- Hawking radiation, 1109
- HCN. *See* Hydrogen cyanide (HCN)
- HD 140283 (HD 140243), 61
- HE 0107–5240, 64, 65, 67, 70, 71, 101, 102
- HE 1327–2326, 67, 70, 71, 80, 101–103
- Heat capacity, 512, 517, 518
- Heating rate, 566, 579
- HeI lines, 262
- HeI stars, 262
- Helium, 1104, 1109, 1110
- Henderson, T., 1096
- Hercules, 1096, 1104
- Herschel, J., 244
- Herschel, W., 1096
- Hertzsprung-Russell diagram, 66
- HES. *See* Hamburg/ESO survey (HES)
- HESS, 791, 793, 796, 797, 799–803, 807, 811–813, 819, 823, 824
- H₂ formation rate, 539–541
- H I, 1099
- H I clouds
- cool, 550, 551, 564, 565, 570, 572, 577–579, 582, 584
- field, 552, 553, 558, 570, 572, 575, 577–581, 583, 584
- thermal, 552, 556, 557, 560, 572, 575–580
- H I column density, 552
- High-resolution mirror assembly (HRMA), 251, 260
- High-resolution spectroscopy, 67, 108–110
- High-velocity clouds, 1103
- history, 592, 615, 636
- High-velocity stars, 3, 5
- H I-line emission, 991, 996, 1029
- H I self-absorption (HISA), 560, 564–565
- HK Survey, 63, 98
- Holmberg, E., 1100–1103, 1105
- Holmberg radius, 1111
- Horizontal branch morphology, 5, 15
- Hot dark matter, 1114
- Hoyle-Lyttleton theory, 263
- HRMA. *See* High-resolution mirror assembly (HRMA)
- Hubble radius, 1111
- Hubble's constant, 1099, 1104, 1108, 1110, 1112, 1116
- Hubble's law, 1096
- Huygens, C., 1096, 1115
- Hyades, 833, 851, 853
- Hydrodynamic and magnetohydrodynamic simulations, 265
- Hydrodynamic infall, 260
- Hydrodynamics simulation, 264, 265
- Hydrogenated amorphous carbon (HAC), 504, 513, 514
- Hydrogen cyanide (HCN), 249, 251
- Hypothesis testing, 125–126
- I**
- Ice, 513, 515, 541–542
- IGIMF, 125, 126, 128, 154, 160, 187, 191, 199, 205, 212, 224–231, 234
- IGM. *See* Intergalactic medium (IGM)
- Imaging atmospheric Cherenkov telescopes, 792
- IMF. *See* Initial mass function (IMF)
- Inflation, 1111, 1112, 1115
- Infrared, 500, 502, 503, 505, 509, 511, 513, 514, 516–517, 522, 523, 528, 540, 545, 546, 1098, 1110

- Infrared astronomical satellite (IRAS), 615
 Infrared flares, 266
 Initial mass function (IMF), 7, 118–234
 discontinuity, 159, 160, 185–188, 198
 of galactic field, 119–121, 123, 128, 142, 143,
 154, 158–162, 170, 172, 178, 179, 187,
 191, 193, 194, 197, 208, 233
 of primaries, 166, 174, 180, 190–191
 of systems, 190–191
 variation, 119, 121, 125, 127, 144, 195, 204,
 209–224, 233
 Inner halo, 5, 6
 Instability, 932–933, 937–946, 949–951, 953, 954,
 958, 965, 976
 Interacting galaxies, 687, 699–705
 radio emission, 700–703
 Intercloud medium, 535–537
 Intergalactic medium (IGM), 645, 700, 701, 710, 711
 magnetic fields, 645, 700, 701, 710, 711
 Intergalactic stars, 1100
 Intermediate-mass black hole, 883, 905, 908,
 911–912
 Intermediate population II, 8
 Intermediate-velocity clouds, 590
 International Astronomical Union, 1094, 1105
 Interstellar clouds, 551, 552, 561
 Interstellar depletion, 506–507
 Interstellar dust
 bubbles, 465, 468, 471, 474, 479–482, 489
 EGOs/YSOs, 476–478, 481
 extinction distribution, 467
 infrared dark clouds, 474–476
 massive star formation regions, 468, 470, 477,
 482–485
 mid-infrared disk images, 468
 MIR extinction/reddening, 468
 M17 MIR, radio, x-ray structure, 482–485
 PAH emission, 468–472
 triggered star formation, 479–482
 Interstellar extinction, 501–504, 509, 511, 513
 Interstellar magnetic field, 579, 584, 799, 823
 Interstellar matter, 987, 988, 993, 997, 999, 1009,
 1012, 1017, 1019–1021, 1030
 Interstellar medium (ISM), 256–258, 262, 263, 267,
 268, 644, 647, 650, 652, 653, 657, 675,
 679, 704, 712, 790, 791, 823
 magnetic fields, 644, 647, 650, 653, 657, 675,
 679, 712
 Interstellar molecules, 539–545
 Interstellar scattering, 267, 268
 Interstellar turbulence, 581
 Intracluster light, 1100
 Intrinsic luminosity, 244
 Intrinsic size, 257, 260
 Inverse Compton scattering, 265, 266
 Ionization potential, 525, 526, 538
 Ionized HVCs, 599–600
 Ionized hydrogen, 245
 IR emission, 501, 504–505, 509, 511–514, 518–523,
 540, 543
 features, 505
 models, 522–523
 spectrum, 504
 IRAS. *See* Infrared astronomical satellite (IRAS)
 Iron-peak elements, 32–34
 Irregular galaxies
 magnetic fields, 680, 693, 694
 radio emission, 694
 IRS 7, 250
 IRS 16, 249, 250, 261, 263
 IRS 16 and IRS 13 complexes, 261
 IRS 13 complex, 261
 IRS 16 complex, 249, 261
 IRS 16/IRS 13, 262
 ISM. *See* Interstellar medium (ISM)
 Isothermal sphere, 884, 886–888, 917
 IV Arch, 595, 605, 610, 611, 616, 629, 634
- J**
 James Webb Space Telescope, 101
 Jeans, J.H., 1096, 1098, 1099, 1102, 1105
 Jeans mass, 123, 124, 187, 201–203, 206, 209,
 211, 213
 Jeans models, 1067–1074, 1079
 Jets, 645, 675, 689, 704–707
 magnetic fields, 645, 705
- K**
 Kaluza-Klein particles, 1113
 Kapteyn, J.C., 1096–1099, 1102
 Kassim, N.E., 245
 3 K background, 1104
 Kepler SNR, 804, 805
 Kinematics, 988, 993, 996, 997, 1003, 1021, 1025,
 1030, 1032
 King model, 884, 888–891, 912
 K-magnitudes, 260
 K-type giants, 262
- L**
 LAB survey, 592, 594, 595, 605, 612
 Lamb shift, 1112
 Large and Small Magellanic clouds, 502, 503
 Large Magellanic Cloud (LMC), 514, 834–846, 848,
 850, 852, 856, 859–864, 1113
 Large-scale anisotropies, 779
 Large-scale clustering, 1112
 Large-scale homogeneity, 1112
 Larmor radius, 532

Late-type stars, 262
 Leading arm, 616, 631
 Leaky box, 734–736, 739
 Leptonic models, 804, 807, 811, 825
 Lick observatory, 1100
 Life cycle, 500, 501, 527–538, 545
 Light bending, 266, 267
 Lithium abundance, 58, 79, 80, 101
 LMC. *See* Large Magellanic Cloud (LMC)
 Local Group, 590, 602, 988, 1010, 1021, 1025–1029, 1035, 1099, 1102
 Lop-sided, 938, 940–941, 953, 976
 Lorentz factor, 259, 265
 Lorentz oscillator, 509, 510
 Lowest observable metallicity, 68
 Luminosity function (LF) of stars, 169
 Lutz-Kelker, 837, 843
 Lutz-Kelker-type bias, 837

M

M-0.02-0.07, 256
 M31, 2, 7, 8, 12–13, 17, 597, 599, 602, 625–627, 632, 1097, 1099, 1100, 1102, 1103, 1109
 M33, 1101, 1103
 M 101, 622–623, 626
 MACHO, 1113
 Magellanic cloud, 502, 503, 514
 Magellanic stream, 590, 593, 595, 601, 608, 610, 611, 616, 618, 627, 630–632, 635
 MAGIC, 791, 796, 817, 818, 823
 Magnetic fields, 1107, 1115
 dissipation, 264
 evolution, 644, 645, 658, 661, 694, 708, 709, 711, 712
 origin, 646, 648, 653–657, 676, 680, 698, 708, 711
 strength, 644–648, 650–652, 654, 656, 658, 659, 664, 667, 669, 670, 674–680, 689–691, 693, 694, 696, 698–700, 705, 707–711
 structure, 644, 645, 651, 657, 661, 662, 664, 666, 667, 670–672, 674, 680–690, 693, 700, 708, 710–712
 Magnetohydrodynamic (MHD) simulation, 264, 265
 Main sequence, 256, 261, 262
 Malmquist-type bias, 837, 867, 870
 Mandl, R.W., 1107
 Maser spots, 255
 2MASS, 603, 612
 Mass distribution, 987–1035
 Mass flow rate, 630
 Mass function, 260
 Mass loss, 883, 895, 897–898, 905, 910, 912, 914–917
 Mass loss rates, 530, 531
 Mass–luminosity relation of stars, 130, 159, 165

Mass segregation, 155, 171, 173, 174, 176, 195, 196, 204, 205, 214, 884, 898, 900–902, 908, 910–912, 915
 Mass spectrum, 598
 Mass-to-luminosity ratio (M/L), 987, 1003, 1007–1009, 1029–1030, 1032, 1034, 1035
 Mathis-Rumpl-Nordsieck, 510
 Maximum stellar mass, 134, 144–153, 230, 234
 Mayer, T., 1095
 MDF. *See* Metallicity distribution functions (MDF)
 Mercury, 1097
 Mergers captures, 1098
 Metallicity, 2–15
 Metallicity distribution functions (MDF), 61, 63, 67, 72–76, 94
 Metal-poor stars, 5, 10, 57–110
 Milky Way, 4–10, 13, 17, 57, 58, 60, 61, 63, 67, 72, 75–77, 94–96, 100, 105–110, 989, 1019, 1021–1024, 1027, 1028, 1030, 1033, 1034, 1096–1100, 1102, 1103, 1107, 1108, 1113
 galaxy, 4, 10–12, 16, 17
 halo, 75, 81–96, 106
 magnetic fields, 657–672
 radio emission, 658, 664, 670
 Milne, E.A., 1104
 Minihalos, 100, 101
 Minnett, H.C., 244
 Minor merger, 13
 Miras, 844–847, 855–856, 858, 860, 861, 863, 864
 Miras variables, 844–847
 Missing mass, 1097
 Missing-satellite, 106
 Mixed compositional model, 776
 Mixed morphology, 256
 MK system, 851, 852, 865, 870
 $m_{\text{max}}-M_{\text{cl}}$ relation, 131–135, 148–154, 156, 204, 208, 209, 212, 221, 225, 227, 231, 233, 234
 M-0.02-0.07 molecular cloud, 256
 Mode, 933, 938–941, 943–954, 965, 966, 976
 Model atmospheres, 26, 42, 51
 MODified Newtonian Dynamics (MOND), 1115
 Molecular and gas clouds, 245
 Molecular clouds, 247, 255, 256, 791, 814, 822–825
 Molecular gas, 249
 Molecular hydrogen, 503, 539–541, 601, 616, 1098, 1099
 MOND. *See* MODified Newtonian Dynamics (MOND)
 Monolithic collapse, 5, 6
 Monopoles, 1109, 1112
 Monte-Carlo codes, 892, 894–895
 Morrison, P., 1107
 Most metal-poor stars, 68, 95
 Moving cluster, 833
 MRN grain size distribution, 512
 MRN model, 510

Multiphoton events, 520, 523
 Multi-wavelength observations, 258, 259
 Murchison meteorite, 506

N

Na I, 607
 National Radio Astronomy Observatory,
 248–251, 256
 N-body codes, 832, 883, 892, 893
 Near-infrared, 260, 267
 Neptune, 1093, 1094, 1107
 Neutral medium, 551, 577
 Neutrinos, 1104, 1109, 1112–1116
 Neutron capture elements, 34–35
 Neutron half-life, 1110
 Neutron matter, 1110
 Neutron stars, 252, 253
 NGC 891, 623, 624, 626, 634
 NGC 3115, 1103
 NGC 4258, 837, 838, 859
 NGC 6946, 624–626
 Non-baryonic DM, 1108
 Non-gravitational forces, 1105, 1106
 Non-local thermodynamic equilibrium (Non LTE),
 70–71, 77, 78, 80, 82, 86, 108
 Non-Newtonian gravity, 1104
 Nonthermal synchrotron, 247
 Novae, 528, 529, 850
 Nuclear wind, 256
 Nucleation, 527, 530, 531
 Nucleo-chronometry
 Nucleosynthesis, 23, 29, 33, 34, 51
 actinide boost, 98, 99
 r-process, 87, 90–94
 s-process, 87–90

O

OB stars, 262, 835, 851–852, 856, 857, 861, 865,
 867, 868
 Oder-Neiss, 1094
 OH masers, 255–257
 One-dimensional atmospheres, 62, 69–71
 One-dimensional model atmosphere analyses, 62,
 69–71
 Oort constants, 834, 847, 857, 868
 Oort, J.H., 1097–1100, 1102–1104
 Oort limit, 1099
 Open clusters, 348–389, 837, 847, 851,
 853–854, 859
 Optical depths, 502, 504, 507, 517, 521, 604,
 606–608, 1107
 Optimal sampling, 124, 125, 131–135, 147–152,
 194–197, 225, 226, 233

Orbit, 925, 928–931, 935–939, 942, 946–948, 953,
 958–962, 965, 966, 970, 973, 975
 Outer halo, 5, 6, 9
 Outflows, 250, 253, 255, 263
 O VI, 590, 596, 597, 601, 617–622, 628, 630
 Oxides, 507, 514, 530, 531
 Oxygen, 605–607, 609, 617, 622
 Oxygen abundances, 28–31, 39

P

PAHs. *See* Polycyclic aromatic hydrocarbons (PAHs)
 Pair-instability SNe, 101
 Palomar Observatory Sky Survey, 1105
 Paradigm shift, 1094
 Parallax, 1095, 1096
 Paranal, 261
 Parkes, 592, 605
 Particle physics, 1070, 1075–1081
 Pattern speed, 929, 930, 934, 939, 944–946, 948, 949,
 952, 958, 961, 963–964, 967
 π^0 decay gamma rays, 795, 803, 817, 819, 821
 Peebles, P.J.E., 1107, 1108, 1110, 1112
 Perihelion advance, 1097
 PeVatrons, 792, 819–822, 825
 Phase space, 927, 950, 960, 966, 973
 Phase transitions, 1109
 Philolaus, 1093, 1094
 Photoelectric, 500, 523–527, 532, 538–540
 Photoelectric efficiency, 538–540
 Photoelectric heating, 538–540
 Photoelectric rates, 525–526
 Photoelectric yield, 525
 Piazzini, G., 1094
 Piddington, H., 244
 Planetary nebulae, 1100, 1105
 Pleiades, 852, 853
 Plummer model, 883, 885–886, 908, 909
 Point sources, 782–783
 Point-source spectrum, 252
 Point-spread function, 260, 267
 Poisson equations, 1098
 Polarization
 optical, 643–646, 649, 652, 657–658, 672–674
 radio synchrotron, 674
 sub-millimeter, 644, 646, 708
 Polycyclic aromatic hydrocarbons (PAHs), 501–543
 clusters, 505, 512, 514, 539
 molecules, 500, 501, 505, 512, 513, 519, 520, 523
 spectrum, 519–523
 Population I, 2–7, 10
 Population II, 2–15
 Population III, 7, 60, 68, 100–103
 Population III stars, 60, 82, 96, 102, 103
 Positrons, 736, 738, 747, 750–751, 757, 783
 Potential, 1002, 1007, 1011–1015, 1028, 1033

- Pre-stellar matter, 1105
 Primary cosmic-ray nucleus, 752
 Primordial black holes, 1105, 1109
 Pristine presolar SIC, 506
 Processing, 500, 503, 532–535, 543, 545
 Procyon, 1093, 1094
 Procyon B, 1099
 Proper motions, 1095, 1096, 1102
 Protostars, 262
 Pseudobulges, 10, 11, 16
 Pulsars, 664–668, 670–672, 677, 678, 710, 711, 1109, 1115
 Pulsation parallaxes, 833–834, 837, 841–843, 859
 Pygmy galaxies, 1100
- Q**
- QSOs, 1108, 1109
 Quasars, 1108
 Quintessence, 1112
- R**
- Radial velocity, 1096, 1100, 1102
 Radiative phase, 532
 Radioactive elements, 1110, 1116
 Radio image, 245, 248, 251
 Radio-infrared correlation, 678–680
 Radio interferometers, 553
 Radio morphology, 247–250
 Radio surveys, 663
 Ratio of total to selective extinction, 502
 Rayleigh limit, 508
 Ray-tracing, 267
 Recombination, 1108
 Red clump, 847–849, 856, 862, 864
 Red giant(s), 250, 261
 Red supergiants, 528, 545
 Reduced parallax method, 838, 869–871
 Refractive scintillation, 258
 Reheating, 1112
 Relative abundances, 61, 75–78, 84–86, 94, 96, 97
 Relative ages, Age_{NORM} for the Galactic globular clusters, 105
 Resonance, 929–933, 935, 941, 942, 947, 949, 950, 952, 954, 960–963, 966–968, 970, 975
 Reverse shock, 806–811, 814, 817, 819
 ROSAT, 250, 252
 ROSAT all-sky survey, 252
 ROSAT PSPC, 253
 Rotation curves, 551, 566–570, 987–1035, 1103, 1105, 1109–1110
 21 cm, 1109
r-process-enhanced metal-poor stars, 93, 98
r-process-enhanced stars, 92, 98–100, 110
 RR Lyraes, 833, 839, 841–844, 849, 854, 856, 858, 860–864
 stars, 603
 variables, 244, 841–843
 Rutile, 537
 RX J1347–1145, 1114
 RXJ1713.7–3946, 796, 799–807
- S**
- Sagittarius, 244, 245, 247
 Sagittarius A, 244, 245, 247, 252, 255
 Sagittarius A*, 245, 247, 248, 250, 251, 255–266, 268
 Sagittarius A* cluster, 261, 262
 Sagittarius A complex, 245, 247
 Sagittarius A East, 245, 247, 248, 255, 256, 260
 Sagittarius A West, 245, 247–251, 260, 263
 Sagittarius B1, 247
 Sagittarius B2, 247, 857, 858
 Sagittarius D, 247
 Sagittarius dwarf spheroidal galaxy, 12
 Satellite galaxy, 12, 13
 Savary, F., 1096
 SBBN. *See* Standard big bang nucleosynthesis (SBBN)
 Scale height, 558, 571–573, 575, 584
 Scattered, 500, 503, 508
 Scattered/fluoresced photon field, 255
 Scattering, 930, 946, 949, 950, 969–974
 albedo, 503
 efficiency, 508
 Schwarzschild, K., 1096–1097
 Schwarzschild, M., 1095, 1101, 1103
 Schwarzschild radius, 258, 265, 266, 268, 1095
 Scintillators, 734, 742, 743, 758, 759, 761, 773
 Sct stars, 862, 864
 SDSS. *See* Sloan digital sky survey (SDSS)
 Second parameter, 5, 6, 15
 Secular evolution, 968–974
 Sedov-Taylor phase, 532
 Selected areas, 1098
 Seyfert, 247
 Seyfert galaxies, 1108
 Shadow, 266, 268, 269
 Shapley, H., 244, 1097, 1099, 1100
 Sheared sheet, 935, 941, 942, 947, 972
 Shocks, 255, 263, 264, 500, 503, 506, 513, 527, 532–537, 540, 545, 1115
 Silicate, 503–505, 507–511, 513–517, 522, 527–531, 537, 538, 542
 Simple stellar population, 2, 15
 Simulation, 262–265, 268, 927, 937–944, 946–950, 952–959, 962, 963, 965, 967, 968, 971, 972, 974, 975
 Single dish telescopes, 258

- Sirius, 1093, 1094, 1096, 1107
- Size distribution, 500, 507, 509–511, 517, 523, 533, 538, 539, 541
- Sizes, 501, 508–510, 512, 521, 522, 532, 539
- SKA. *See* Square kilometer array (SKA)
- Sky mapper, 109
- Sloan digital sky survey (SDSS), 603, 612
- Small grains, 505, 509, 511, 512, 520, 523, 525, 532–534, 538, 539, 541, 546
- Small-scale inhomogeneities, 1112
- Small-scale structure, 601, 605, 606, 610–614
- SN 1006, 796–799, 804, 814
- SN 1987, 861
- SN 1987A, 514, 528, 530, 531
- SNR 0.9+0.1, 247
- SNRs. *See* Supernova remnants (SNRs)
- SN yields, 102, 103
 - calculations, 81
- Solar apex, 1096
- Solar neighborhood, 1099
- Solar system, 1096, 1098, 1106, 1115
- Sources of interstellar dust, 528–529
- Space telescope imaging spectrograph (STIS), 617
- Spatial power spectrum, 580–582
- Special relativity, 1095
- Spectroscopy, 63–67, 108–110
 - binary, 1096
 - parallaxes, 850–851
- Spectrum synthesis, 99
- Spheroidal halo, 1110
- Spinel, 515, 537
- Spinning black hole, 266
- Spin temperature, 558, 575, 583
- Spiral arms, 644, 668, 669, 673, 675, 677, 678, 681, 682, 693, 712
 - magnetic fields, 644, 668, 669, 673, 675, 677, 682, 693, 712
- Spiral galaxies, 1102, 1111
 - magnetic fields, 675, 680–690
 - radio emission, 694
- Spirals, 925, 930, 931, 934–936, 938, 940–950, 952, 956, 962–966, 970–974, 976, 1103, 1110
- Spite plateau, 79, 80, 101
- Sputtering, 501, 532–535, 537
- Square kilometer array (SKA), 583, 674, 680, 689, 708–712
- Stability criterion/condition, 933, 939
- Standard big bang nucleosynthesis (SBBN), 79–81
- Standard candle, 244
- Star-burst, 262
- Star clusters, 118, 120–122, 124–131, 142–145, 147, 148, 150, 151, 153–157, 160–162, 164, 169–178, 185, 187, 194, 195, 198, 202, 204–206, 208, 210, 214, 215, 219, 220, 222, 225–227, 231, 233, 234, 348, 350–352, 357, 361–365, 370, 388, 1105, 1109, 1114
- Star counts
 - color-magnitude relation
 - globular clusters, 407, 408
 - nearby stars, 407–409
 - fundamental equation, 404–407, 413
 - Hess diagram, 420–426
 - luminosity function
 - globular clusters, 397, 406, 407, 421
 - nearby stars, 405–407, 421
 - major starcount program
 - Basel program, 414
 - Besancon program, 414
 - SDSS program, 414, 415
- Stardust, 501, 505–506, 514, 516, 531, 536, 537
- Stars
 - asymptotic giant branch (AGB) stars, 487
 - luminous blue variables/WR stars, 488–489
 - planetary nebulae, 487–488
 - super nova remnants (SNRs), 489
- Statistical parallaxes, 832–833, 843, 851, 853
- Steady State, 1104, 1110
- Stellar abundances, 364, 375, 376
- Stellar collisions, 884, 892, 905–906, 912
- Stellar cusp, 263
- Stellar dynamics, 1099
- Stellar evaporation, 915, 917, 919
- Stellar evolution, 349, 352, 358, 359, 362–365, 375
- Stellar halo, 5, 8, 9, 12–14
- Stellar kinematics, 40
- Stellar metallicity, 26, 27
- Stellar nucleosynthesis, 352, 364–365
- Stellar populations, 2–17, 23, 27, 29–36, 40, 43, 48, 50–52, 274, 278, 279, 296, 325, 330, 332, 337
- Sterile neutrinos, 1076, 1078, 1079, 1114, 1115
- Sticking, 524, 536, 540, 541
 - coefficient, 524, 536, 541
 - probability, 540, 541
- STIS. *See* Space telescope imaging spectrograph (STIS)
- Strings, 1109
- Strömberg, G., 1097
- Stromgren photometry, 6, 7, 11
- Strong lensing, 1107
- Structure formation, 1111, 1114, 1115
- Struve, W., 1096
- Subdwarfs, 3, 9, 603
- Sulfur, 605–607, 609
- Superclustering, 1106
- Supergiants, 851–8523
- Supermassive black hole, 245, 246, 248, 250, 255–263
- Supernovae, 514, 516, 527–529, 531, 532, 534, 543, 546, 1095, 1105, 1113, 1116

Supernova remnants (SNRs), 245, 247, 248, 253,
790–825

 ejecta, 528, 531, 808, 809, 812, 817

Super partners, 1113

Superposition model, 754, 769

Surface brightness method, 833

Surface gravity, 25, 28, 36, 51

Swing amplifier/amplification, 925–977

S XV, 253

Symmetry breaking, 1109

Synchrotron, 245, 247, 265, 266

Synchrotron emission, 255, 258, 259, 265, 266, 643,
646–650, 659, 660, 674, 678, 680, 700,
706–712

T

100'' Telescope, 1099

Telescope resolution, 267

Temperature, 501, 505, 512, 514, 516–524, 527, 530,
531, 533, 535, 536, 538, 540,
541, 543

Temperature distribution function, 519–522

Temperature fluctuations, 512, 517–518, 520

Textures, 1109

Thermal brehmsstrahlung, 1106

Thermal plasma, 253, 257

Thermal radio emission, 645, 658, 679

Thermodynamic condensation sequence, 529, 531

Thick disks, 7–9, 16, 30, 31, 33, 35–40, 42–45, 47, 48,
50–52

 evolution

 chemical evolution, 407, 429, 431, 433, 434,
 439, 440

 gravitational settling, 396, 433, 435

 viscous evolution, 436–438

 formation, 431–441

 double exponential structure, 434–439

 observational constraints

 abundance ratio, 430–431

 age, 430

 color distribution, 397, 421–425

 external galaxies, 405, 432, 437

 metallicity distribution, 404, 405, 410, 412,
 425, 426, 429, 439, 441

 normalization, 397, 398, 401–403, 407, 410,
 412, 415, 417–419, 425, 429, 432

 radial scale length, 418–419

 rotational lag, 427, 428

 velocity dispersion, 417, 427, 428, 432, 435,
 436, 439, 441

 vertical scale height, 417–418

Thin disks, 8, 9, 11–13, 16, 17, 29–31, 33, 35–40, 42,
43, 50–52

Thirty meter telescope, 109

Three-armed spiral, 248

Three-dimensional, 70

Three-dimensional model atmospheres, 70, 71

Tidal capture, 904, 906–907, 909

Tidal disruption, 266

Tidal forces, 262

Tidal interactions, 627

Tidal streams, 589, 630–632

Tidal truncation, 915

Timescales, 610, 612–614, 635

Tip of the red giant branch (TRGB), 848–850, 862

Topological defects, 1114

Transition radiation detectors, 742, 743

Transition region, 770, 773–776, 778, 784

Transition temperature gas, 617, 619, 621, 622, 628

Transport equation, 735

TRGB. *See* Tip of the red giant branch (TRGB)

TRGB(L), 863

TRGB(K), 860, 863

Trigonometrical parallaxes, 831–832, 837, 838, 843,
845, 846, 850, 853, 860

T-Tauri star, 530

Tully-Fisher relation, 864, 867, 868, 871

Turbulence, 612, 613, 1115

2,175Å bumps, 502, 503, 509, 513, 514, 523

2,175Å feature, 502, 503, 509, 513, 514, 523

Two-body interactions, 262

Two-body relaxation, 882, 884, 895, 897–900,
914, 918

Tycho SNR, 796, 804, 805

Type II Cepheids, 833, 835, 839–841, 850, 860, 864

U

Uhuru satellite, 1106

Ultra-faint dwarf galaxies, 75, 77, 95, 108, 1043,

1047–1049, 1051, 1056, 1058, 1059, 1062,

1063, 1065, 1066, 1070, 1073

Ultraviolet nonstellar continuum, 256

Unenlightened stars, 1093, 1094

Universe, 57–110

Unseen mass, 1097

UV excess, 4, 5

UV lines, 604, 606

V

Vacuum field, 1112

van Maanen, A., 1099

Vaporization, 527, 533, 534

Vela Jr. SNR, 796, 804

Velocity dispersion, 1051–1072, 1074, 1075, 1078

VERITAS, 791, 796, 817, 818, 823

Very large array (VLA), 245, 247, 248, 255, 259, 260,
605, 623

 radio image, 245, 248

 telescope, 245

Very long baseline interferometry (VLBI), 256, 257, 268
 Very small grains, 505, 511, 512, 520, 523, 525, 538, 539, 546
 Villa Elisa telescope, 592
 Violent relaxation, 895–897
 Virgo, 1099, 1104
 Virgo cluster, 687, 701–703, 1100
 Virial equilibrium, 601
 Virial theorem, 1106, 1108
 VLA. *See* Very large array (VLA)
 VLBI. *See* Very long baseline interferometry (VLBI)
 Voids, 1114
 von Zach, F.X., 1094

W

Warm dark matter, 1078, 1114
 Warm neutral medium, 551
 Warp, 950, 954–957
 Wave theory of light, 1094
 ω Centauri, 15, 75, 76, 94
 WC Wolf-Rayet, 528
 Weakly interacting massive particles (WIMPs), 1076, 1079, 1080, 1113
 Westerbork synthesis radio telescope (WSRT), 605, 612
 Western Arc, 260
 White dwarfs, 252, 253, 1093, 1094, 1097, 1099, 1103, 1113

Wilkinson microwave anisotropy probe (WMAP), 79–80, 100, 1108, 1113, 1116
 WIMPs. *See* Weakly interacting massive particles (WIMPs)
 Winds, 1105
 Wind–wind collisions, 259, 263
 WMAP. *See* Wilkinson microwave anisotropy probe (WMAP)
 Wolf-Rayet stars, 262
 WSRT. *See* Westerbork synthesis radio telescope (WSRT)

X

Xi Uma, 1096
 XMM-Newton, 266, 793, 797, 821
 X-ray and radio telescopes, 247
 X-ray background, 251
 X-ray emission, 250, 252, 255, 259, 260, 263, 1104
 X-ray intensity map, 255
 X-ray interferometry, 269
 X-ray map, 259, 260
 X-ray morphology, 250–256
 X-ray objects, 255
 X-ray plume, 251, 253

Z

Zeeman effect, 643, 644, 649, 653, 664, 668–669, 672–675, 708
 Zeldovich, Y.B., 1107–1109, 1111, 1112