David Holcman *Editor*

Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology

Springer

Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology

David Holcman
Editor

Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology

Springer

Editor
David Holcman
Institute for Biology
École Normale Supérieure
Applied Mathematics
and Computational Biology
Paris, France

Churchill College
University of Cambridge
Cambridge, UK

ISBN 978-3-319-62626-0 ISBN 978-3-319-62627-7 (eBook)
DOI 10.1007/978-3-319-62627-7

Library of Congress Control Number: 2017952704

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

A little physics can make a long way in explaining, quantifying and predicting biological phenomena, especially when propelled by rational, deductive and efficient methods of mathematical analysis and computations.

Preface

In the past 30 years, progress in molecular and cellular biology has greatly benefited from the diversity of experimental methods. At the same time, the amount of data generated has been plethoric, posing real challenges for mathematics, theoretical physics, computer science and more areas. These challenges are extracting hidden features from large and high-dimensional data sets and generating fast multiscale simulations. To accomplish this program, various tools were further developed: fast stochastic simulations to simulate stochastic chemical reactions have been built on the Gillespie method. Other developments include the derivation of biophysical modelling or reducing the complexity of high-dimensional stochastic processes by projection into low-dimensional space, where the analysis is possible. Finally, deriving asymptotic formula has revived asymptotic analysis of partial differential equations, because they usually represent the new physical laws and clarify the role of singular parameters.

The convergence of these interests and techniques has engulfed mathematical biology into a new area at the intersection of statistical physics, applied mathematics, applied probability, computer science, biophysics and cell biology. This new field does not simply provide new tools to extract features from data or to simulate large amount of particles, but aims to contribute in quantifying and explaining the function of a cell or its subcellular components from the molecular organization (from nano- to micro- and higher scales).

Our young community has recently been challenged by producing modelling, analysis, effective and fast methods of computing, based on coarse-graining or analysis of the model equations. Examples of biological processes include diffusion in microdomains, calcium dynamics, gene regulation, chromatin organization and modification, signal transduction and molecular signalling, coagulation-fragmentation of proteins and cellular transport, but also synaptic formation and plasticity, cellular communication, neuronal network organization, early patterning during development and many more.

Molecular and cellular biology processes are inherently stochastic, which is the main driving force of many biological functions such as during ionic channels or synaptic transmission. Stochastic and rare events are at the basis of signal

transduction, but also facilitate phenotypic diversity of cellular populations and even drive mutation during evolution. At the scale of a single cell, stochasticity becomes relevant due to low copy numbers of biological molecules, such as mRNA or transcription factors, that take part in biochemical reactions driving cellular processes. When trying to describe such biological processes, the traditional mean-field or coarse-grained deterministic models are often inadequate, exactly because of these low copy numbers. But stochastic models are necessary to account for small particle numbers (intrinsic noise) and extrinsic noise sources. The complexity of these models depends crucially on whether the biochemical reactions are diffusion-limited or reaction-limited. In the latter case, processes are described by adopting the framework of Markov jumps and stochastic differential equations (chemical master and Fokker-Planck equations), while in the former it is possible to adopt the framework of stochastic reaction-diffusion models, including reaction-diffusion master equation, partial differential equations and particle-based Brownian dynamics simulations.

This book is divided into four mains parts. The first describes stochastic and master chemical reactions with low copy numbers. The method involves dimensional reduction. The second concerns the theory and method of random simulations using stochastic processes for motion, but also chemical reactions. The third is dedicated to asymptotic analysis used to explore the parameter space. The fourth explores diffusion processes and stochastic modelling in cell biology. Several examples of cell biological systems are treated here such as the model of axonal growth and analysis of photoresponse with an emphasis in the multiscale chemical reactions for the signal transduction inside rod photoreceptors. At a cell population level, several stochastic models are introduced about the mitochondrial heterogeneity across network configurations and genetic heterogeneity within cells and between generations. Finally, using birth, death, immigration and local dispersal of individual's model, some empirical stochastic equations and observables are introduced to study spatial pattern organization.

The chapters are written for a large audience of mathematicians, physicists, computational biologists and computer scientists interested in studying stochastic and numerical methods and physical modelling for cellular processes. This collective effort originates from a 6-month programme that took place at the Newton Institute in Cambridge in 2016, organized by R. Erban, K. Zygalakis, S. Isaacson and myself about, "Stochastic Dynamical Systems in Biology: Numerical Methods and Applications" (https://www.newton.ac.uk/event/sdb). We thank the Newton Institute and its director John Tolland and the Simons Foundation for making this programme possible that resulted in the present book.

Cambridge, UK David Holcman
July 2016

Contents

Contributors

Assyr Abdulle ANMC, Institut de Mathmatiques, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Sandro Azaele Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds, UK

Kevin Burrage ARC Centre of Excellence for Mathematical and Statistical Frontiers, School of Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, QLD, Australia

Pamela Burrage ARC Centre of Excellence for Mathematical and Statistical Frontiers, School of Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, QLD, Australia

Mario Castro Department of Applied Mathematics, University of Leeds, Leeds, UK

Grupo Intedisciplinar de Sistemas Complejos (GISC), Universidad Pontificia Comillas, Madrid, Spain

P.H. Constantino Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN, USA

Stefan Engblom Division of Scientific Computing, Department of Information Technology, Uppsala University, Uppsala, Sweden

Radek Erban Mathematical Institute, University of Oxford, Oxford, UK

Ramon Grima School of Biological Sciences, University of Edinburgh, Edinburgh, UK

Andreas Hellander Division of Scientific Computing, Department of Information Technology, Uppsala University, Uppsala, Sweden

E. Hingant Departmento de Matemática, Universidad del Bío-Bío, Concepción, Chile

Hanne Hoitzing Imperial College London, London, UK

David Holcman Institute for Biology École Normale Supérieure, Applied Mathematics and Computational Biology, Paris, France

Churchill College, University of Cambridge, Cambridge, UK

N. Hoze Institut für Integrative Biologie, ETH, Zürich, Switzerland

Gregory Jameson Biophysics Graduate Program, Ohio State University, Columbus, OH, USA

Iain G. Johnston School of Biosciences, University of Birmingham, Birmingham, UK

Nick S. Jones Imperial College London, London, UK

Y.N. Kaznessis Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN, USA

Andre Leier Department of Genetics and Informatics Institute, University of Alabama at Birmingham, School of Medicine, Birmingham, AL, USA

Per Lötstedt Division of Scientific Computing, Department of Information Technology, Uppsala University, Uppsala, Sweden

Grant Lythe Department of Applied Mathematics, University of Leeds, Leeds, UK

Tatiana Marquez-Lago Department of Genetics and Informatics Institute, University of Alabama at Birmingham, School of Medicine, Birmingham, AL, USA

Carmen Molina-París Department of Applied Mathematics, University of Leeds, Leeds, UK

Fabio Peruzzo Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds, UK

Tomislav Plesa Mathematical Institute, University of Oxford, Oxford, UK

Jürgen Reingruber INSERM U1024; Applied Mathematics and Computational Biology, IBENS, Ecole Normale Supérieure, Paris, France

Stephen Smith School of Biological Sciences, University of Edinburgh, Edinburgh, UK

K. Spiliopoulos Department of Mathematics and Statistics, Boston University, Boston, MA, USA

Tomáš Vejchodský Institute of Mathematics, Czech Academy of Sciences, Žitná, Praha, Czech Republic

M. Vlysidis Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN, USA

Chuan Xue Department of Mathematics, Ohio State University, Columbus, OH, USA

R. Yvinec CR2 INRA, UMR85 Physiologie de la Reproduction et des Comportements, F-37380 Nouzilly, France

Part I
Stochastic Chemical Reactions

Test Models for Statistical Inference: Two-Dimensional Reaction Systems Displaying Limit Cycle Bifurcations and Bistability

Tomislav Plesa, Tomáš Vejchodský, and Radek Erban

1 Introduction

Given noisy time-series, it may be of practical importance to infer possible biological mechanisms underlying the time-series [1]. Mathematically, such statistical inferences correspond to an inverse problem, consisting of mapping given noisy time-series to compatible reaction networks. One way to formulate the inverse problem is as follows. Firstly, one obtains deterministic kinetic ordinary-differential equations (ODEs) compatible with the stochastic time-series. And secondly, suitable reaction networks may then be induced from the obtained kinetic ODEs [2, 3]. The inverse problem is generally ill-posed [2, 3], as more than one suitable reaction networks may be obtained. In order to make a progress in solving the inverse problem, it is useful to impose further constraints on the kinetic ODEs. A particular set of constraints on the kinetic ODEs may be obtained by determining the types of the deterministic attractors which are 'hidden' in the noisy time-series [1]. This may be a challenging task, especially when cycles (oscillations) are observed in the time-series. The observed cycles may be present in both the deterministic and stochastic models (also known at the stochastic level as *noisy deterministic cycles*), or they may be present only in the stochastic model (also known as *quasi-cycles*, or noise-induced oscillations). Noisy deterministic cycles may arise directly from the autonomous kinetic ODEs, or via the time-periodic terms present in the nonautonomous kinetic ODEs. Quasi-cycles may arise from the intrinsic or extrinsic noise, and have been shown to exist near deterministic stable foci, and stable

T. Plesa (✉) • R. Erban
Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK
e-mail: plesa@maths.ox.ac.uk

T. Vejchodský
Institute of Mathematics, Czech Academy of Sciences, Žitná 25, Praha 1, 115 67, Czech Republic

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_1

nodes [4]. For two-species reaction systems, quasi-cycles can be further classified into those that are unconditionally noise-dependent (but dependent on the reaction rate coefficients), and those that are conditionally noise-dependent [4]. Thus, a cycle detected in a noisy time-series may at the deterministic level generally correspond to a stable limit cycle, a stable focus, or a stable node.

In order to detect and classify cycles in noisy time-series, several statistical methods have been suggested [1, 5]. In [1], analysis of the covariance as a function of the time-delay, spectral analysis (the Fourier transform of the covariance function), and analysis of the shape of the stationary probability mass function have been suggested. Let us note that reaction systems of the Lotka-Volterra (x-factorable [2]) type are used as test models in [1], and that conditionally noise-dependent quasi-cycles, which can arise near a stable node, and which can induce oscillations in only a subset of species [4], have not been discussed. In addition to the aforementioned statistical methods developed for analyzing noisy time-series, methods for (locally) studying the underlying stochastic processes near the deterministic attractors/bifurcations have also been developed [4, 6–11].

Statistical and analytical methods for studying cycles in stochastic reaction kinetics have often been focused on deterministically monostable systems which undergo a local bifurcation near a critical (equilibrium) point, known as the supercritical Hopf bifurcation. We suspect this is partially due to simplicity of the bifurcation, and partially due to the fact that it is difficult to find two-species reaction systems, which are more amenable to mathematical analysis, undergoing more complicated bifurcations and displaying bistability involving limit cycles. Nevertheless, kinetic ODEs arising from biological applications may exhibit more complicated bifurcations and multistabilites [12–14]. Thus, it is of importance to test the available methods on simpler test models that display some of the complexities found in the applications.

In this paper, we construct two reaction systems that are two-dimensional (i.e. they only include two chemical species) and induce cubic kinetic equations, first of which undergoes a global bifurcation known as a convex supercritical homoclinic bifurcation, and which displays bistability involving a critical point and a limit cycle (which we call mixed bistability). The second system undergoes a local bifurcation known as a multiple limit cycle bifurcation, and displays bistability involving two limit cycles (which we call bicyclicity). Aside from finding an application as test models for statistical inference and analysis in biology, to our knowledge, the constructed systems are also the first examples of two-dimensional reaction systems displaying the aforementioned types of bifurcations and bistabilities. Let us note that reaction systems with dimensions higher than two, displaying the homoclinic bifurcation, as well as bistabilities involving two limit cycles, have been reported in applications [12–14].

The reaction network corresponding to the first system is given by

$$
\begin{aligned}
r_1 &: \quad \varnothing \xrightarrow{k_1} s_1, & r_7 &: \quad \varnothing \xrightarrow{k_7} s_2,\\
r_2 &: \quad s_1 \xrightarrow{k_2} 2s_1, & r_8 &: \quad s_2 \xrightarrow{k_8} \varnothing,\\
r_3 &: \quad 2s_1 \xrightarrow{k_3} 3s_1, & r_9 &: \quad s_1 + s_2 \xrightarrow{k_9} s_1 + 2s_2,\\
r_4 &: \quad s_1 + s_2 \xrightarrow{k_4} s_2, & r_{10} &: \quad 2s_2 \xrightarrow{k_{10}} 3s_2,\\
r_5 &: \quad 2s_1 + s_2 \xrightarrow{k_5} s_1 + s_2, & r_{11} &: \quad 3s_2 \xrightarrow{k_{11}} 2s_2,\\
r_6 &: \quad s_1 + 2s_2 \xrightarrow{k_6} 2s_1 + 2s_2, & & \qquad (1)
\end{aligned}
$$

where the two species s_1 and s_2 react according to the eleven reactions $r_1, r_2, \ldots, r_{11}$ under mass-action kinetics, with the reaction rate coefficients denoted $k_1, k_2, \ldots, k_{11}$, and with $\varnothing$ being the zero-species [2]. A particular choice of the (dimensionless) reaction rate coefficients is given by

$$
\begin{aligned}
&k_1 = 0.01, \quad k_2 = 0.9, \quad k_3 = 1.55, \quad k_4 = 2.6, \quad k_5 = 1.2, \quad k_6 = 1.5,\\
&k_7 = 0.01, \quad k_8 = 3.6, \quad k_9 = 1, \quad k_{10} = 2.4, \quad k_{11} = 0.8, \qquad (2)
\end{aligned}
$$

while more general conditions on these parameters are derived later as Eqs. (10) and (11).

The reaction network corresponding to the second system includes two species s_1 and s_2 which are subject to the following fourteen chemical reactions $r_1, r_2, \ldots, r_{14}$:

$$
\begin{aligned}
r_1 &: \quad \varnothing \xrightarrow{k_1} s_1, & r_8 &: \quad \varnothing \xrightarrow{k_8} s_2,\\
r_2 &: \quad s_1 \xrightarrow{k_2} \varnothing, & r_9 &: \quad s_2 \xrightarrow{k_9} 2s_2,\\
r_3 &: \quad 2s_1 \xrightarrow{k_3} 3s_1, & r_{10} &: \quad s_1 + s_2 \xrightarrow{k_{10}} s_1,\\
r_4 &: \quad s_1 + s_2 \xrightarrow{k_4} 2s_1 + s_2, & r_{11} &: \quad 2s_2 \xrightarrow{k_{11}} 3s_2,\\
r_5 &: \quad 3s_1 \xrightarrow{k_5} 4s_1, & r_{12} &: \quad 2s_1 + s_2 \xrightarrow{k_{12}} 2s_1 + 2s_2,\\
r_6 &: \quad 2s_1 + s_2 \xrightarrow{k_6} s_1 + s_2, & r_{13} &: \quad s_1 + 2s_2 \xrightarrow{k_{13}} s_1 + s_2,\\
r_7 &: \quad s_1 + 2s_2 \xrightarrow{k_7} 2s_2, & r_{14} &: \quad 3s_2 \xrightarrow{k_{14}} 2s_2, \qquad (3)
\end{aligned}
$$

where $k_1, k_2, \ldots, k_{14}$ are the corresponding reaction rate coefficients. A particular choice of the (dimensionless) reaction coefficients is given by[1]

[1]Let us note that the limit cycles corresponding to (3) are *highly* sensitive to changes in the parameters (4). Thus, during numerical simulations, parameters (4) should *not* be rounded-off. One can also design bicyclic systems which are less parameter sensitive, see Appendix 2.

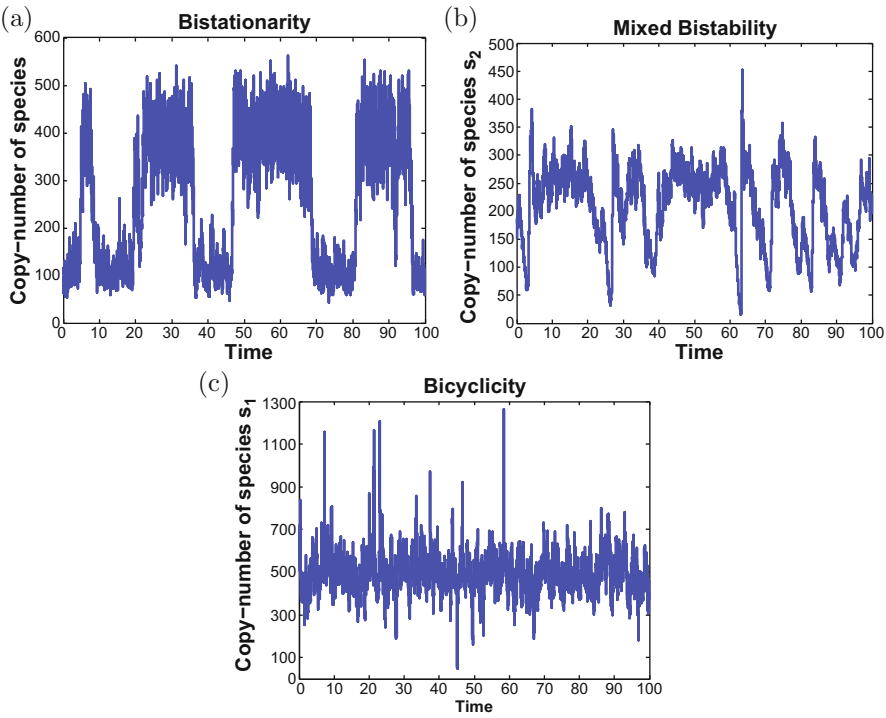


Fig. 1 Panels (**a**), (**b**), and (**c**) show representative sample paths generated using the Gillespie stochastic simulation algorithm for the Schlögl system [15] with coefficients as in [6], reaction network (1) with coefficients (2) and reactor volume $V = 100$, and reaction network (3) with coefficients (4) and $V = 0.5$, respectively. At the deterministic level, the phase planes of (1) and (3) are shown in Fig. 2. The deterministic and stochastic time-series, as well as the probability distributions, are shown in Figs. 3 and 4. At the deterministic level, a critical point and a limit cycle are 'hidden' in (**b**), while two limit cycles are 'hidden' in (**c**)

$$k_1 = 2 \times 10^{-7}, \quad k_2 = 19.987880407, \quad k_3 = 0.019944378,$$
$$k_4 = 0.02003132232, \quad k_5 = 2.9 \times 10^{-8}, \quad k_6 = 2.000232 \times 10^{-5},$$
$$k_7 = 1.45 \times 10^{-8}, \quad k_8 = 2 \times 10^{-7}, \quad k_9 = 8.38734, \quad k_{10} = 0.038389,$$
$$k_{11} = 0.0215726, \quad k_{12} = 2 \times 10^{-5}, \quad k_{13} = 1.571 \times 10^{-6}, \quad k_{14} = 10^{-5}, \quad (4)$$

while the general conditions on these parameters are given later as Eqs. (13) and (14).

In Fig. 1, we display a representative noisy-time series generated using the Gillespie stochastic algorithm, in Fig. 1a for the one-dimensional cubic Schlögl system [15], which deterministically displays two stable critical points (bistationarity [3]), in Fig. 1b for the reaction network (1) with coefficients (2), which deterministically displays a stable critical point and a stable limit cycle (mixed

bistability), and in Fig. 1c for the reaction network (3) with coefficients (4), which deterministically displays two stable limit cycles (bicyclicity). Several statistical challenges arise. For example, is it possible to infer that the upper attractor in Fig. 1b is a deterministic critical point, while the lower a noisy limit cycle? Is it possible to detect one/both noisy limit cycles in Fig. 1c? The answer to the second question is complicated by the fact that the two deterministic limit cycles in Fig. 1c are relatively close to each other.

The rest of the paper is organized as follows. In Sect. 2, we outline properties of the planar quadratic ODE systems, concentrating on cycles, cycle bifurcations and multistability. There are two reasons for focusing on the planar quadratic systems: firstly, the phase plane theory for such systems is well-developed [16, 17], with a variety of concrete examples with interesting phase plane configurations [18–20]. Secondly, an arbitrary planar quadratic ODE system can always be mapped to a kinetic one using only an affine transformation—a special property not shared with cubic (nor even linear) planar systems [21]. This, together with the available nonlinear kinetic transformations which increase the polynomial degree of an ODE system by one [2], implies that we may map a general planar quadratic system to at most cubic planar kinetic system, which may still be biologically or chemically relevant. In Sect. 3, we present the two planar cubic test models which induce reaction networks (1) and (3), and which are constructed starting from suitable planar quadratic ODE systems. We also compare the deterministic and stochastic solutions of the constructed reaction networks, and highlight the observed qualitative differences. Finally, in Sect. 4, we provide a summary of the paper.

2 Properties of Two-Dimensional Second-Degree Polynomial ODEs: Cycles, Cycle Bifurcations and Multistability

Let us consider the two-dimensional second-degree autonomous polynomial ODEs

$$\begin{aligned}\frac{\mathrm{d}x_1}{\mathrm{d}t} &= \mathcal{P}_1(x_1, x_2; \mathbf{k}) = k_1 + k_2 x_1 + k_3 x_2 + k_4 x_1^2 + k_5 x_1 x_2 + k_6 x_2^2,\\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= \mathcal{P}_2(x_1, x_2; \mathbf{k}) = k_7 + k_8 x_1 + k_9 x_2 + k_{10} x_1^2 + k_{11} x_1 x_2 + k_{12} x_2^2,\end{aligned} \tag{5}$$

where $\mathcal{P}_i(\cdot, \cdot; \mathbf{k}) : \mathbb{R}^2 \to \mathbb{R}$, $i \in \{1, 2\}$, are the second-degree two-variable polynomial functions, and $\mathbf{k} = (k_1, k_2, \ldots, k_{12}) \in \mathbb{R}^{12}$ is the vector of the corresponding coefficients. We assume that $\mathcal{P}_1$ and $\mathcal{P}_2$ are relatively prime and at least one is of second-degree. We allow coefficients $\mathbf{k}$ to be parameter-dependent, $\mathbf{k} = \mathbf{k}(\mathbf{p})$, with $\mathbf{p} \in \mathbb{R}^q$, $q \geq 0$.

Let us consider two additional properties which system (5) may satisfy:

(I) Coefficients $k_1,\ k_3,\ k_6,\ k_7,\ k_8,\ k_{10} \geq 0$, i.e. $\mathcal{P}_1$ and $\mathcal{P}_2$ are so-called kinetic functions (for a rigorous definition see [2]).

(II) The species concentrations $x_1 = x_1(t)$ and $x_2 = x_2(t)$ are uniformly bounded in time for $t \geq 0$ in the nonnegative orthant $\mathbb{R}^2_{\geq}$, except possibly for initial conditions located on a finite number of one-dimensional subsets of $\mathbb{R}^2_{\geq}$, where infinite-time blow-ups are allowed.

The subset of Eqs. (5) satisfying properties (I)–(II) are referred to as the *deterministic kinetic equations* bounded in $\mathbb{R}^2_{\geq}$, and denoted

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= \mathcal{K}_1\big(x_1, x_2; \mathbf{k}(\mathbf{p})\big), \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= \mathcal{K}_2\big(x_1, x_2; \mathbf{k}(\mathbf{p})\big). \end{aligned} \tag{6}$$

In what follows, we discuss only the biologically/chemically relevant solutions of (6), i.e. the solutions in the nonnnegative quadrant $\mathbb{R}^2_{\geq}$. We now summarize some of the definitions and results regarding cycles, cycle bifurcations and multistability (referred to as the so-called exotic phenomena in the biological context [3]) for systems (5) and (6). Let us note that most of the results have been shown to hold only for the more general system (5), and may not necessarily hold for the more restricted system (6).

Critical Points A (finite) critical point $(x_1^*(\mathbf{k}), x_2^*(\mathbf{k}))$ of system (5) is a solution of the polynomial system $\mathcal{P}_1(x_1^*, x_2^*; \mathbf{k}) = 0, \mathcal{P}_2(x_1^*, x_2^*; \mathbf{k}) = 0$. Critical points are the time-independent solutions of (5).

Cycles Cycles of (5) are closed orbits in the phase plane which are not critical points. They can be isolated (limit cycles, and separatrix cycles) or nonisolated (a one-parameter continuous family of cycles). Limit cycles are the periodic solutions of (5). A homoclinic separatrix cycle consists of a homoclinic orbit and a critical point of saddle type, with the orbit connecting the saddle to itself. On the other hand, a heteroclinic separatrix cycle consists of two heteroclinic orbits, and two critical points, with the orbits connecting the two critical points [22]. Limit cycles of (6) correspond to biological clocks, which play an important role in fundamental biological processes, such as the cell cycle, the glycolytic cycle and circadian rhythms [23–25].

Cycle Bifurcations Variations of coefficients $\mathbf{k}$ in (5) may lead to changes in the topology of the phase plane (e.g. a change may occur in the number of invariant sets or their stability, shape of their region of attraction or their relative position). Variation of $\mathbf{k}(\mathbf{p})$ in (6) may be interpreted as a variation of the reaction rate coefficients $\mathbf{k}$ due to changes in the reactor (environment) parameters $\mathbf{p}$, such as the pressure or temperature. If the variation causes the system to become topologically nonequivalent, such a parameter is called a bifurcation parameter, and at the parameter value where the topological nonequivalence occurs, a bifurcation is said to take place [22, 26]. Bifurcations in the deterministic kinetic equations have been reported in applications [12, 23–25, 27, 28].

Bifurcations of limit cycles of (5) can be classified into three categories: (i) the Andronov-Hopf bifurcation, where a limit cycle is created from a critical point of focus or center type, (ii) the separatrix cycle bifurcation, where a limit cycle is created from a separatrix cycle, and (iii) the multiple limit cycle bifurcation, where a limit cycle is created from a limit cycle of multiplicity greater than one [16, 22]. Let us note that the maximum multiplicity of a multiple focus of (5) is three, so that at most three local limit cycles can be created under appropriate perturbations [29]. Bifurcations (i) and (iii) are examples of local bifurcations, occurring in a neighborhood of a critical point or a limit cycle, while bifurcations (ii) are examples of global bifurcations, occuring near a separatrix cycle. The following global bifurcations may occur in (5): convex homoclinic bifurcations (defined in, e.g., [30]), saddle–saddle (heteroclinic) bifurcations, and the saddle-node (heteroclinic) bifurcations on an invariant cycle. However, concave homoclinic bifurcations, double convex, and double concave homoclinic bifurcations, presented in, e.g., [30], cannot occur in (5) as a consequence of basic properties of planar quadratic ODEs [31, 32].

A necessary condition for the existence of a limit cycle in (6) is that $k_4 > 0$ or $k_{12} > 0$ [2, 3]. This implies that the induced reaction network must contain at least one autocatalytic reaction of the form $2s_i \to ns_i + ms_j$, with $n \geq 3$, $m \geq 0$, and $i, j \in \{1, 2\}$. In the literature, system (6) has been shown to display the following limit cycle bifurcations: Andronov-Hopf bifurcations, saddle-node on an invariant cycle, and multiple limit cycle bifurcations [21, 33, 34]. Let us note that some of the reaction systems constructed in [21, 33, 34] (e.g. displaying double Andronov-Hopf bifurcation, and a saddle–saddle bifurcation) are described by ODEs of the form (6), but with solutions which are generally not bounded in $\mathbb{R}^2_{\geq}$.

Multistability System (5) is said to display multistability if the total number of the underlying stable critical points and stable limit cycles is greater than one, for a fixed $\mathbf{k}$. Multistability in (6) corresponds to biological switches, which may be classified into reversible or irreversible [27, 35, 36]. The former switches play an important role in reversible biological processes (e.g. metabolic pathways dynamics, and reversible differentiation), while the latter in irreversible biological processes (e.g. developmental transitions, and apoptosis).

Multistability can be mathematically classified into *pure multistability*, involving attractors of only the same type (either only stable critical points, or only stable limit cycles), and *mixed multistability*, involving at least one stable critical point, and at least one stable limit cycle. Pure multistability involving only critical points is called *multistationarity* [3], while we call pure multistability involving only limit cycles *multicyclicity*. Mixed bistability, and bicyclicity, can be further classified into concentric and nonconcentric. Concentric mixed bistability (resp. bicyclicity) occurs when the stable limit cycle encloses the stable critical point (resp. when the first stable limit cycle encloses the second stable limit cycle), while nonconcentric when this is not the case. Let us note that, for a fixed kinetic ODE system (6), multistationarity at some parameter values $\mathbf{k}$, is neither necessary, nor sufficient, for cycles at some (possibly other) parameter values $\mathbf{k}'$ [37].

We now prove that (5) can have at most three coexisting stable critical points, i.e. (5) can be at most *tristationary*.

Lemma 2.1 *The maximum number of coexisting stable critical points in two-dimensional relatively prime second-degree polynomial ODE systems* (5)*, with fixed coefficients* **k***, is three.*

Proof Let us assume system (5) has four, the maximum number, of real finite critical points. Then, using an appropriate centroaffine (linear) transformation [31, 32], system (5) can be mapped to

$$\begin{aligned}\frac{\mathrm{d}x_1}{\mathrm{d}t} &= a_1x_1(x_1-1)+b_1x_2(x_2-1)+c_1x_1x_2,\\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= a_2x_1(x_1-1)+b_2x_2(x_2-1)+c_2x_1x_2,\end{aligned} \tag{7}$$

which is topologically equivalent to (5), with the critical points located at $A=(0,0)$, $B=(1,0)$, $C=(0,1)$ and $D=(\alpha,\beta)$, with $\alpha\neq 0$, $\beta\neq 0$, $\alpha+\beta\neq 1$, and the coefficients c_1, c_2 given by

$$\begin{aligned}c_1 &= -\frac{\alpha-1}{\beta}a_1-\frac{\beta-1}{\alpha}b_1,\\ c_2 &= -\frac{\alpha-1}{\beta}a_2-\frac{\beta-1}{\alpha}b_2.\end{aligned}$$

The trace and determinant of the Jacobian matrix of (7), denoted τ and δ, respectively, evaluated at the four critical points, A, B, C, D, are given by:

$$\begin{aligned}
\tau_A &= -(a_1+b_2), & \delta_A &= a_1b_2-a_2b_1,\\
\tau_B &= a_1-a_2\frac{(\alpha-1)}{\beta}-b_2\frac{(\alpha+\beta-1)}{\alpha}, & \delta_B &= -\frac{\alpha+\beta-1}{\alpha}\delta_A,\\
\tau_C &= b_2-a_1\frac{(\alpha+\beta-1)}{\beta}-b_1\frac{(\beta-1)}{\alpha}, & \delta_C &= -\frac{\alpha+\beta-1}{\beta}\delta_A,\\
\tau_D &= \alpha a_1+\beta b_2-a_2\frac{\alpha(\alpha-1)}{\beta}-b_1\frac{\beta(\beta-1)}{\alpha}, & \delta_D &= (\alpha+\beta-1)\delta_A.
\end{aligned} \tag{8}$$

System (7) may have three stable critical points if and only if the quadrilateral $ABCD$, formed by the critical points, is nonconvex, and the only saddle critical point is the one located at the interior vertex of the quadrilateral [31, 32]. This is the case when $\alpha>0$, $\beta>0$, $\alpha+\beta<1$, and $\delta_A>0$, in which case A, B, and C are nonsaddle critical points, while D is a saddle. Imposing also the conditions $\tau_A<0$, $\tau_B<0$, $\tau_C<0$, ensuring that A, B, and C are stable, a solution of the resulting system of algebraic inequalities is given by $a_1=1$, $b_1=-1$, $a_2=1$, $0<\alpha<1/2\left((1+2\beta)-\sqrt{1+8\beta^2}\right)$, $-1<b_2<\alpha(-\alpha+\beta+1)/(\beta(\alpha+\beta-1))$. □

Let us note that if (7) is kinetic, then it cannot have three stable critical points. More precisely, requiring $b_1 \geq 0$, $a_2 \geq 0$, and $d_A > 0$ and $\tau_A < 0$ in (8) implies $a_1 > 0$ and $b_2 > 0$, which further implies $\tau_B > 0$, so that B is unstable. More generally, the authors have not found a tristationary system (6) in the literature (and we conjecture it does not exist). On the other hand, bistationary systems (6) do exist (in fact, even one-dimensional cubic bounded kinetic systems may be bistationary, e.g. the Schlögl model [15], see the time-series shown in Fig. 1a).

The maximum number of stable limit cycles in (5) is two, i.e. (5) can be at most *bicyclic*. Furthermore, system (5) may also display *mixed tristability*, involving one stable critical point, and two stable limit cycles. This follows from the fact that the maximum number of limit cycles in (5) is four, in the unique configuration (3, 1), a fact only recently proved in [17], solving the second part of Hilbert's 16th problem for the quadratic case. If the solutions of (5) are required to be bounded in the whole $\mathbb{R}^2$, system (5) was conjectured to have at most two limit cycles [22, 38], and hence have at most one stable limit cycle. It remains an open problem if the maximum number of limit cycles in the nonnegative orthant of (6) is four or less (we conjecture it is less than four), and if (6) may be bicyclic. Due to the fact that (6) is (I) kinetic (and, hence, nonnegative), and (II) appropriately bounded in $\mathbb{R}^2_{\geq}$, additional restrictions are imposed on the boundary of $\mathbb{R}^2_{\geq}$, and on the critical points at infinity, complicating the construction of systems (6) displaying multistability involving limit cycles. Some results regarding multistability have been obtained in [21]: system (6) displaying concentric mixed bistability has been constructed. The system contains two limit cycles in the nonnegative orthant, and therefore does not exceed the conjectured bound on the number of limit cycles in the bounded quadratic systems [22, 38]. While a kinetic system of the form (6) displaying concentric bicyclicity has been obtained in [21], the system is not bounded in $\mathbb{R}^2_{\geq}$.

3 Test Models: Construction and Simulations

In this section, our aim is to construct two-dimensional kinetic ODEs bounded in $\mathbb{R}^2_{\geq}$, which display a nonconcentric bistability. As highlighted in the previous section, it may be a difficult task to obtain such systems with at most quadratic terms, i.e. in the form (6). To make a progress, in this section, we allow the two-dimensional kinetic ODEs to contain cubic terms, and we construct two systems. The first system displays a convex homoclinic bifurcation, and mixed bistability, and is obtained by modifying a system from [2] using the results from Appendix 1. The second system displays a multiple limit cycle bifurcation, and bicyclicity. To construct the second system, we use an existing system of the form (5), which forms a one-parameter family of uniformly rotated vector fields [22, 39], and which displays bicyclicity and multiple limit cycle bifurcation [40]. We use kinetic transformations from [2] to map this system, which is of the form (5), to a kinetic one, which is of the form (6). We then use the results from Appendix 1 to map the system of the form (6) to a suitable cubic two-dimensional kinetic system. We

also fine-tune the polynomial coefficients in the kinetic ODEs in such a way that sizes of the two stable limit cycles differ by maximally one order of magnitude (a task that can pose challenges [18]). As differences may be observed between the deterministic and stochastic solutions for parameters at which a deterministic bifurcation occurs [6], we investigate the constructed models for such observations. Let us note that an alternative static (i.e. not dynamic) approach for reaction system construction, using only the chemical reaction network theory or kinetic logic, provides only conditions for stability of critical points, but no information about the phase plane structures [41], and is, thus, insufficient for construction of the systems presented in this paper.

3.1 System 1: Homoclinic Bifurcation and Mixed Bistability

Consider the following deterministic kinetic equations

$$\begin{aligned}\frac{\mathrm{d}x_1}{\mathrm{d}t} &= k_1 + x_1\left(k_2 + k_3x_1 - k_4x_2 - k_5x_1x_2 + k_6x_2^2\right),\\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= k_7 + x_2\left(-k_8 + k_9x_1 + k_{10}x_2 - k_{11}x_2^2\right),\end{aligned} \tag{9}$$

with the coefficients $\mathbf{k} = \mathbf{k}(a, \boldsymbol{\mathcal{T}}, \alpha, \varepsilon)$ given by

$$\begin{aligned} k_1 &= \varepsilon, & k_7 &= \varepsilon,\\ k_2 &= \frac{1}{2}\left|\left(3\left(\mathcal{T}_2 - \frac{2}{3}\right)(a\mathcal{T}_1 + \mathcal{T}_2) - 2\alpha\mathcal{T}_1\right)\right|, & k_8 &= |-\mathcal{T}_1 + a\mathcal{T}_2(\mathcal{T}_2 - 1)|,\\ k_3 &= \left|-\frac{3}{2}a\left(\mathcal{T}_2 - \frac{2}{3}\right) + \alpha\right|, & k_9 &= 1,\\ k_4 &= \left|1 - \frac{3}{2}(a\mathcal{T}_1 + 2\mathcal{T}_2)\right|, & k_{10} &= \left|2a\left(\mathcal{T}_2 - \frac{1}{2}\right)\right|,\\ k_5 &= \left|\frac{3}{2}a\right|, & k_{11} &= |a|,\\ k_6 &= \frac{3}{2}, \end{aligned} \tag{10}$$

where $|\cdot|$ denotes the absolute value, and with parameters a, α, ε, $\mathcal{T}_1$, and $\mathcal{T}_2$ satisfying

$$\begin{aligned} &a \in (-1, 0), \quad |\alpha| \ll 1, \quad 1 \ll \varepsilon \leq 0,\\ &\mathcal{T}_1 > \frac{2\sqrt{3}}{9}, \quad \mathcal{T}_2 \in \left(\max(1, -a\mathcal{T}_1), \frac{2}{3} + \frac{8}{3}a^{-2}\left(3 - a^2\right)\left(a + 4\mathcal{T}_1\right)\right). \end{aligned} \tag{11}$$

The canonical reaction network [2] induced by system (9) is given by (1).

System (9) is obtained from system [2, Eq. (32)], which is known to display a mixed bistability and a convex supercritical homoclinic bifurcation when $\alpha = 0$, $\varepsilon = 0$. We have modified [2, Eq. (32)] by adding to its right-hand side the ε-term from Definition 1 [i.e. coefficients k_1 and k_7 in (9)], thus preventing the long-term dynamics to be trapped on the phase plane axes. It can be shown, using Theorem 1, that choosing a sufficiently small $\varepsilon > 0$ in (10) does not introduce additional positive critical points in the phase space of (9).

In Fig. 2a, b, we show phase plane diagrams of (9) before and after the bifurcation, respectively, where the critical points of the system are shown as the colored dots (the stable node, saddle, and unstable focus are shown as the green, blue and red dots, respectively), the blue curves are numerically approximated saddle manifolds (which at $\alpha = 0$, $\varepsilon = 0$ form a homoclinic loop [2]), and the purple curve in Fig. 2b is the stable limit cycle that is created from the homoclinic separatrix cycle. Let us note that parameter α, appearing in (10), controls the bifurcation, while parameter a controls the saddle-node separation [2].

In Fig. 3a–b and d–e, we show numerical solutions of the initial value problem for (9) in red, with one initial condition in the region of attraction of the node, while the other near the unstable focus. The blue sample paths are generated by using the Gillespie stochastic simulation algorithm on the induced reaction network (1), initiated near the unstable focus. More precisely, in Fig. 3a, d we show the dynamics before the deterministic bifurcation, when the node is the globally stable critical point for the deterministic model, while in Fig. 3b, e we show the dynamics after the bifurcation, when the deterministic model displays mixed bistability. On the other hand, the stochastic model displays relatively frequent stochastic switching in Fig. 3a, b, when the saddle-node separation is relatively small. Let us emphasize that the stochastic switching is observed even before the deterministic bifurcation. In Fig. 3d, e, when the saddle-node separation is relatively large, the stochastic switching is significantly less common, and the stochastic system in the state-space is more likely located near the stable node. Thus, in Fig. 3d, e, the stochastic system is less affected by the bifurcation than the deterministic system, and, in fact, behaves more like the deterministic system before the bifurcation. This is also confirmed in Fig. 3c, f, where we display the x_2-marginal stationary probability mass functions (PMFs) for the smaller and larger saddle-node separations, respectively, which were obtained by numerically solving the chemical master equation (CME) [42, 43] corresponding to network (1). Let us note that, by sufficiently increasing the saddle-node separation, the left peak in the PMF from Fig. 3f, corresponding to the deterministic limit cycle, becomes nearly zero and difficult to detect.

In [44], we present an algorithm which structurally modifies a given reaction network under mass-action kinetics, in such a way that the deterministic dynamics is preserved, while the stochastic dynamics is modified in a controllable state-dependent manner. We apply the algorithm on reaction network (1), for parameter values similar as in Fig. 3d–f, to make the underlying PMF bimodal, so that the underlying sample paths display stochastic switching between the two deterministic attractors. Furthermore, we also make the PMF unimodal, and concentrated around the deterministic limit cycle, so that the underlying sample paths remain near the deterministic limit cycle. Meanwhile, we preserve the deterministic dynamics induced by (9).

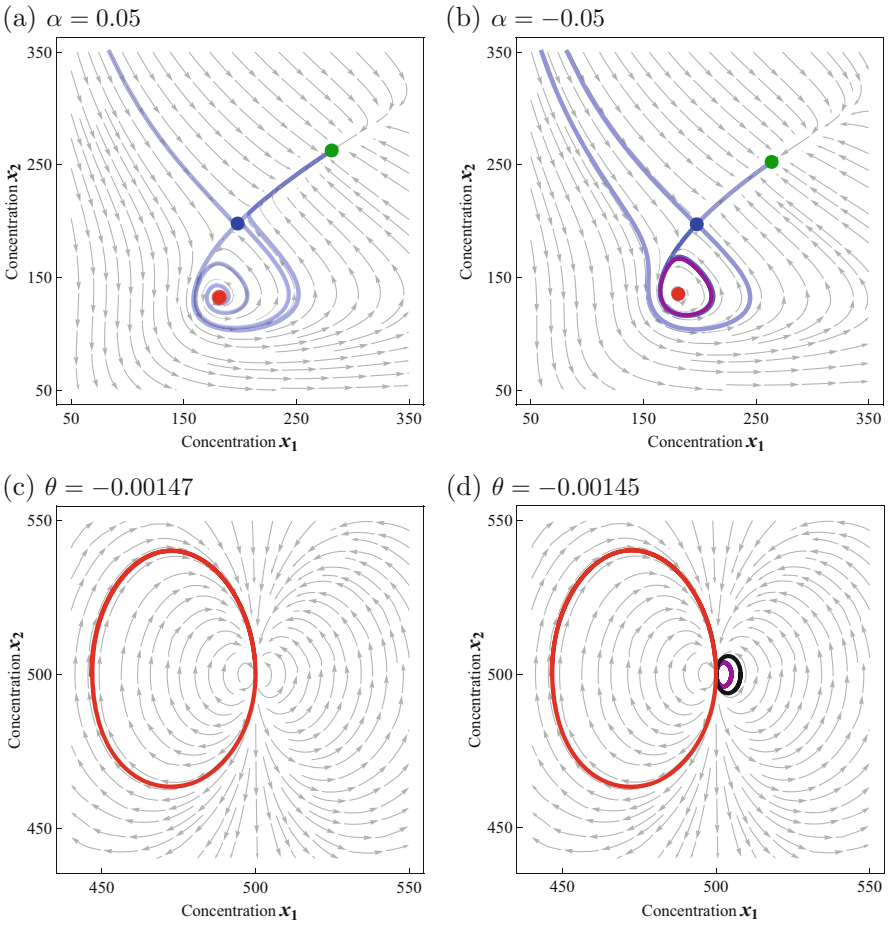


Fig. 2 (**a**)–(**b**) Phase plane diagrams of system (9) before and after the homoclinic bifurcation. The stable node, saddle, and unstable focus are represented as the *green, blue and red dots*, respectively, the vector field as *gray arrows*, numerically approximated saddle manifolds as *blue trajectories*, and the *purple curve* in panel (**b**) is the stable limit cycle. The parameters appearing in (10), and satisfying (11), are fixed to $a = -0.8$, $\mathcal{T}_1 = \mathcal{T}_2 = 2$, $\varepsilon = 0.01$, the reactor volume is set to $V = 100$, and the bifurcation parameter α is as shown in the panels. (**c**)–(**d**) Phase plane diagrams of system (12) before and after the multiple limit cycle bifurcation. The stable limit cycles L_1 and L_3 are shown in *purple and red*, respectively, while the unstable limit cycle L_2 is shown in *black*. The parameters appearing in (13), and satisfying (14), are fixed to $a = 1$, $b = -1$, $c = 0.5$, $d = 0.08$, $x_1^* = -3$, $\mathcal{T}_1 = \mathcal{T}_2 = 1000$, $\varepsilon = 0.01$, the reactor volume is set to $V = 0.5$, and the bifurcation parameter θ is as shown in the panels

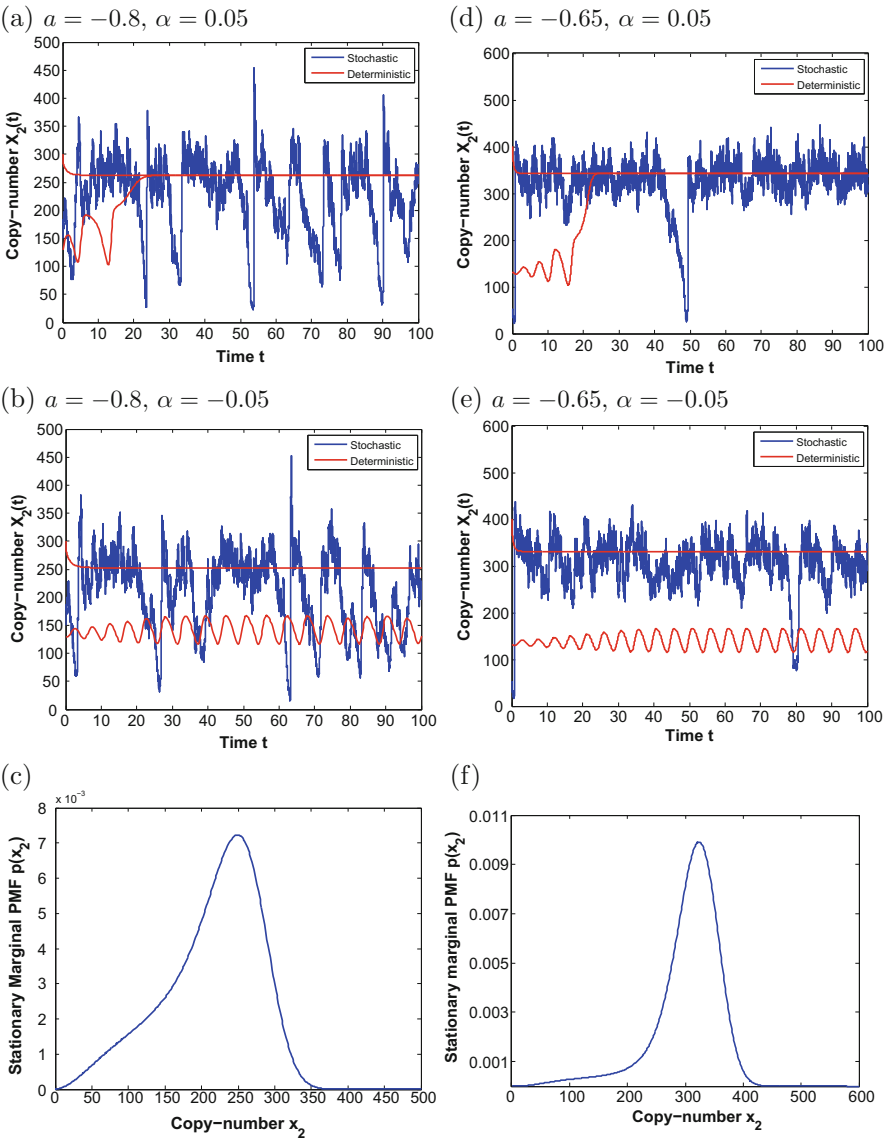


Fig. 3 Numerical solutions of system (9) are shown in *red*. Representative sample paths, generated by the Gillespie stochastic simulation algorithm applied on the corresponding reaction network (1), are shown in *blue*. Probability mass functions (PMFs), obtained by numerically solving the underlying chemical master equation (CME) on the bounded domain $(x_1, x_2) \in [0, 1000] \times [0, 600]$, are also shown in *blue*. (**a**)–(**b**) The cases before and after the homoclinic bifurcation, respectively, for smaller values of a, when the limit cycle and the stable node are closer together. (**d**)–(**e**) The cases before and after the homoclinic bifurcation, respectively, for larger values of a. (**c**) and (**f**) Stationary x_2-marginal PMFs. Parameter values in (**c**) and (**f**) are the same as in (**b**) and (**e**), respectively. One of the deterministic solutions is initiated in the region of attraction of the node, while the other near the focus. The parameters are fixed to $\mathcal{T}_1 = \mathcal{T}_2 = 2$, $\varepsilon = 0.01$, the reactor volume is set to $V = 100$, with a and α as shown in the panels

3.2 System 2: Multiple Limit Cycle Bifurcation and Bicyclicity

Consider the following deterministic kinetic equations

$$\begin{aligned}\frac{\mathrm{d}x_1}{\mathrm{d}t} &= k_1 + x_1\left(-k_2 + k_3x_1 + k_4x_2 + k_5x_1^2 - k_6x_1x_2 - k_7x_2^2\right),\\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= k_8 + x_2\left(k_9 - k_{10}x_1 + k_{11}x_2 + k_{12}x_1^2 - k_{13}x_1x_2 - k_{14}x_2^2\right),\end{aligned} \tag{12}$$

with coefficients $\mathbf{k} = \mathbf{k}(a, b, c, d, x_1^*, \boldsymbol{\mathcal{T}}, \theta, \varepsilon)$ given by

$$\begin{aligned}k_1 &= k_8 = \varepsilon,\\ k_2 &= \left| -a\mathcal{T}_1\mathcal{T}_2\cos(\theta) + \left[(d(\mathcal{T}_1+1) + c\mathcal{T}_2)\mathcal{T}_2 + b(\mathcal{T}_1+1)(\mathcal{T}_1+x_1^*)\right]\sin(\theta)\right|,\\ k_3 &= \left|a\mathcal{T}_2\cos(\theta) - \left[d\mathcal{T}_2 + b(2\mathcal{T}_1 + x_1^* + 1)\right]\sin(\theta)\right|,\\ k_4 &= \left|a\mathcal{T}_1\cos(\theta) - [d(\mathcal{T}_1+1) + 2c\mathcal{T}_2]\sin(\theta)\right|,\\ k_5 &= |b\sin(\theta)|,\\ k_6 &= |-a\cos(\theta) + d\sin(\theta)|,\\ k_7 &= |c\sin(\theta)|,\end{aligned} \tag{13}$$

and if $k_i = \left|f(a, b, c, d, x_1^*, \boldsymbol{\mathcal{T}})\cos(\theta) - g(a, b, c, d, x_1^*, \boldsymbol{\mathcal{T}})\sin(\theta)\right|$, then $k_{i+7} = \left|f(a, b, c, d, x_1^*, \boldsymbol{\mathcal{T}})\sin(\theta) + g(a, b, c, d, x_1^*, \boldsymbol{\mathcal{T}})\cos(\theta)\right|$, $i = 2, 3, \ldots, 7$, and with parameters a, b, c, d, x_1^*, $\mathcal{T}_1$, $\mathcal{T}_2$, θ and ε satisfying

$$\begin{aligned}&0 \le \varepsilon \ll 1, \quad -1 \ll \theta < 0,\\ &b < 0, \quad d > 0, \quad a > -\frac{d^2}{4b}, \quad 0 < c < a + \frac{d^2}{4b}, \quad x_1^* < \frac{d^2}{4bc},\\ &a^3c + b^3(1-x_1^*)^2 \neq 0,\\ &\mathcal{T}_1 > -x_1^*, \quad 0 < \mathcal{T}_2 < -\frac{4abx_1^*}{d^2(x_1^*-1)}(\mathcal{T}_1 + x_1^*),\\ &[d(\mathcal{T}_1+1) + c\mathcal{T}_2]\mathcal{T}_2 + b(\mathcal{T}_1+1)(\mathcal{T}_1+x_1^*) < 0.\end{aligned} \tag{14}$$

The canonical reaction network induced by system (12) is given by (3). In this section, we show that systems (12) and (15) (see below), the latter of which is known to display bicyclicity and a multiple limit cycle bifurcation, are topologically equivalent near the corresponding critical points, provided conditions (14) are satisfied.

In Fig. 2c, d, we show the phase plane diagram of (12) for a particular choice of the parameters satisfying (14), and it can be seen that the system also displays bicyclicity and a multiple limit cycle bifurcation, with Fig. 2c, d showing the cases

before and after the bifurcation, respectively. In Fig. 2c, the only stable invariant set is the limit cycle shown in red, while in Fig. 2d there are two additional limit cycles—a stable one, shown in purple, and an unstable one, shown in black. The purple, black and red limit cycles are denoted in the rest of the paper by L_1, L_2 and L_3, respectively. At the bifurcation point, L_1 and L_2 intersect.

In order to construct (12), let us consider the planar quadratic ODE system [21, 40] given by

$$\begin{aligned}\frac{\mathrm{d}x_1}{\mathrm{d}t} &= \mathcal{Q}_1(x_1,x_2)\cos(\theta) - \mathcal{Q}_2(x_1,x_2)\sin(\theta),\\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= \mathcal{Q}_1(x_1,x_2)\sin(\theta) + \mathcal{Q}_2(x_1,x_2)\cos(\theta),\end{aligned} \tag{15}$$

where

$$\begin{aligned}\mathcal{Q}_1(x_1,x_2) &= -ax_1x_2,\\ \mathcal{Q}_2(x_1,x_2) &= -bx_1^* + b(x_1^*+1)x_1 + dx_2 - bx_1^2 - dx_1x_2 - cx_2^2,\end{aligned} \tag{16}$$

with

$$\begin{aligned}&x_1^* < 0,\ d^2 - 4bcx_1^* < 0,\ d^2 - 4b(c-a) < 0,\\ &\theta d\left(a - b(1-x_1^*)\right) < 0,\ \ \theta bd > 0,\ \ a^3c + b^3(1-x_1^*)^2 \neq 0.\end{aligned} \tag{17}$$

Lemma 3.1 *Consider system (15)–(17), with the real parameter* $\theta \in (-\pi, \pi]$. *Function* $\boldsymbol{\mathcal{P}}(x_1,x_2;\,\theta) = (\mathcal{Q}_1\cos(\theta) - \mathcal{Q}_2\sin(\theta), \mathcal{Q}_1\sin(\theta) + \mathcal{Q}_2\cos(\theta))$ *forms a* one-parameter family of uniformly rotated vector fields *with the rotation parameter* θ, *and the following results hold:*

1. Finite critical points. *System (15) has two critical points in the finite part of the phase plane, located at* $(1,0)$ *and* $(x_1^*,0)$, *both of which are unstable foci when* $|\theta| \ll 1$.
2. Number and distribution of limit cycles. *System (15) has three limit cycles in the configuration* $(2,1)$ *when* $|\theta| \ll 1$. *The focus located at* $(1,0)$ *is surrounded by two positively oriented limit cycles* L_1 *and* L_2, *with the unstable limit cycle* L_2 *enclosing the stable limit cycle* L_1, *while the focus at* $(x_1^*,0)$ *by a single negatively oriented stable limit cycle* L_3.
3. Dependence of the limit cycles on the rotation parameter θ. *There exists a critical value* $\theta = \theta^* < 0$, *at which the limit cycles* L_1 *and* L_2 *intersect in a semistable, positively oriented limit cycle that is stable from the inside, and unstable from the outside. As* θ *is monotonically increased in* $(\theta^*,0)$, *the limit cycles* L_2 *and* L_3 *monotonically expand, while* L_1 *monotonically contracts.*

Proof The statement of the lemma follows from [21, 40], and the theory of one-parameter family of uniformly rotated vector fields [22, 39]. □

In order to map the stable limit cycles of system (15) into the first quadrant, and then map the resulting system to a kinetic one, having no boundary critical points, let us apply a translation transformation $\Psi_{\mathcal{T}}$ [2], $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \mathcal{T}_2) \in \mathbb{R}^2$, followed by a perturbed x-factorable transformation, as defined in Definition 1, on system (15), which results in system (12) with the coefficients (13).

Theorem 3.1 *Consider the ODE systems* (12) *and* (15)*, and assume conditions* (14) *are satisfied. Then* (12) *and* (15) *are locally topologically equivalent in the neighborhood of the corresponding critical points. Furthermore, for sufficiently small* $\varepsilon > 0$*, system* (12) *has exactly one additional critical point in* $\mathbb{R}^2_>$*, which is a saddle located in the neighbourhood of* $(\mathcal{T}_1, 0)$.

Proof Consider the critical point $(1, 0)$ of system (15), which corresponds to the critical point $(\mathcal{T}_1 + 1, \mathcal{T}_2)$ of system (12) when $\varepsilon = 0$. The Jacobian matrices of (15), and (12) with $\varepsilon = 0$, evaluated at $(1, 0)$, and $(\mathcal{T}_1 + 1, \mathcal{T}_2)$, are respectively given by

$$J = \begin{pmatrix} -b(x_1^* - 1)\sin(\theta) & -a\cos(\theta) \\ b(x_1^* - 1)\cos(\theta) & -a\sin(\theta) \end{pmatrix},$$

$$J_{\mathcal{X},\mathcal{T}} = \begin{pmatrix} -b(x_1^* - 1)(\mathcal{T}_1 + 1)\sin(\theta) & -a(\mathcal{T}_1 + 1)\cos(\theta) \\ b(x_1^* - 1)\mathcal{T}_2\cos(\theta) & -a\mathcal{T}_2\sin(\theta) \end{pmatrix}.$$

Condition (ii) of [2, Theorem 3.3] is satisfied, so that the stability of the critical point is preserved under the x-factorable transformation, but condition (iii) is not satisfied. In order for $(\mathcal{T}_1 + 1, \mathcal{T}_2)$ to remain focus under the x-factorable transformation, the discriminant of $J_{\mathcal{X},\mathcal{T}}$ must be negative:

$$(a\mathcal{T}_2 + b(\mathcal{T}_1 + 1)(x_1^* - 1))^2(\sin(\theta))^2 - 4ab(x_1^* - 1)(\mathcal{T}_1 + 1)\mathcal{T}_2 < 0. \tag{18}$$

Let us set $\theta = 0$ in (18), leading to

$$-4ab(x_1^* - 1)(\mathcal{T}_1 + 1)\mathcal{T}_2 < 0. \tag{19}$$

Conditions (18) and (19) are equivalent when $|\theta| \ll 1$, since the sign of the function on the LHS of (18) is a continuous function of θ. From conditions (14) it follows that $ab < 0$, $x_1^* < 0$, and $\mathcal{T}_1, \mathcal{T}_2 > 0$, so that (19) is satisfied. Similar arguments show that the second critical point of (15), located at $(x_1^*, 0)$, is mapped to an unstable focus of (12), if $d > 0$, and if $\mathcal{T}_2$ is bounded as given in (14).

Consider (12) with $\varepsilon = 0$. The boundary critical points are located at $(0, 0)$, $(\mathcal{T}_1, 0)$, and $(0, x_{2,\pm}^*)$, with

$$x_{2,\pm}^* = \frac{1}{2c}\left(d(\mathcal{T}_1 + 1) + 2c\mathcal{T}_2 \pm \sqrt{(\mathcal{T}_1 + 1)(d^2(\mathcal{T}_1 + 1) - 4bc(\mathcal{T}_1 + x_1^*))}\right).$$

Conditions (14) imply that the critical point $(0, 0)$ satisfies $\mathcal{P}_1(0, 0) = -a\mathcal{T}_1\mathcal{T}_2 < 0$, and

$$\mathcal{P}_2(0, 0) = -[d(1 + \mathcal{T}_1) + c\mathcal{T}_2]\mathcal{T}_2 - b(1 + \mathcal{T}_1)(\mathcal{T}_1 + x_1^*) > 0,$$

when $\theta = 0$. When $|\theta| \ll 1$, it then follows from condition (iv) of [2, Theorem 3.3] that the critical point is a saddle, and from Theorem 1, condition (23), that it is mapped outside of $\mathbb{R}^2_{\geq}$ when $\varepsilon \neq 0$. Similar arguments show that, assuming conditions (14) are true, $(\mathcal{T}_1, 0)$ is a saddle that is mapped to $\mathbb{R}^2_{>}$ when $\varepsilon \neq 0$, and that critical points $(0, x^*_{2,\pm})$ are real, $x^*_{2,-} < 0$, and that $(0, x^*_{2,+})$ is a saddle that is mapped outside $\mathbb{R}^2_{\geq}$ when $\varepsilon \neq 0$.

Finally, if conditions (14) are satisfied, so are conditions (17). □

We now consider the kinetic ODEs (12) and the induced reaction network (4) for a particular set of coefficients (13). We also rescale the time according to $t \to 2 \times 10^{-5}\, t$, i.e. we multiply all the coefficients $k_1, \ldots k_{14}$ appearing in (12) by 2×10^{-5}. On this time-scale, we capture dynamical effects relevant for this paper. In Fig. 4a, b we show numerically approximated solutions of the initial value problem for (12) before and after the bifurcation, respectively. In Fig. 4a, we show a solution initiated near the unstable focus, outside the limit cycle L_3. It can be seen that the solution spends some time in a neighborhood of the unstable focus, which is followed by an excursion leading the solution to the stable limit cycle L_3, where it then stays forever. In Fig. 4b, the solutions tend to the limit cycle L_1 or L_3, depending on the initial condition. Let us note that the critical value at which the limit cycles L_1 and L_2 intersect, at the deterministic level, is numerically found to be $\theta^* \approx -0.00146$.

In Fig. 4c, d we show representative sample paths generated by applying the Gillespie stochastic simulation algorithm on the reaction network (3), before and after the bifurcation, respectively. One can notice that the stochastic dynamics does not appear to be significantly influenced by the bifurcation, as opposed to the deterministic dynamics. In Fig. 4c, d, one can notice pulses similar as in Fig. 4a, that are now induced by the intrinsic noise present in the system.

The stationary PMF corresponding to network (3), for parameter values as in Fig. 4c, d, accumulates at the boundary of the state-space (see also the Keizer paradox [45]). While the results from Appendix 1 may be used to prevent a PMF from accumulating at the boundary, one may need a sufficiently large reactor volume. For example, for network (1), the propensity function [43] of reactions r_1 and r_7, for parameter values taken in this paper (i.e. $\varepsilon = 0.01$ in (10), and $V = 100$), takes the value $\varepsilon V = 1$. This is sufficient for the underlying PMF to approximately vanish at the boundary of the state-space, as demonstrated in Fig. 3c and f. On the other hand, for network (3), we take $\varepsilon = 0.01$ in (13), and $V = 0.5$, so that the propensity function of r_1 and r_8 takes the value of only 0.005. As a consequence, the underlying PMF accumulates at the boundary of the state-space. Instead of increasing the reactor volume to prevent this, we instead focus on the so-called quasi-stationary PMF under the condition that the species copy-numbers are positive, $p_>(x, y) \equiv p(x, y | x > 0, y > 0)$. The quasi-stationary PMF describes well the stochastic dynamics of network (3) on the time-scale of interest, presented in Fig. 4c, d. In Fig. 4e, we display an approximate x_1-marginal quasi-stationary PMF $p_>(x_1)$, for the same parameter values as in Fig. 4d. The quasi-stationary PMF $p_>(x_1)$ was obtained by numerically solving the stationary CME corresponding to network (3), on a truncated domain which excludes the boundary of the state-space.

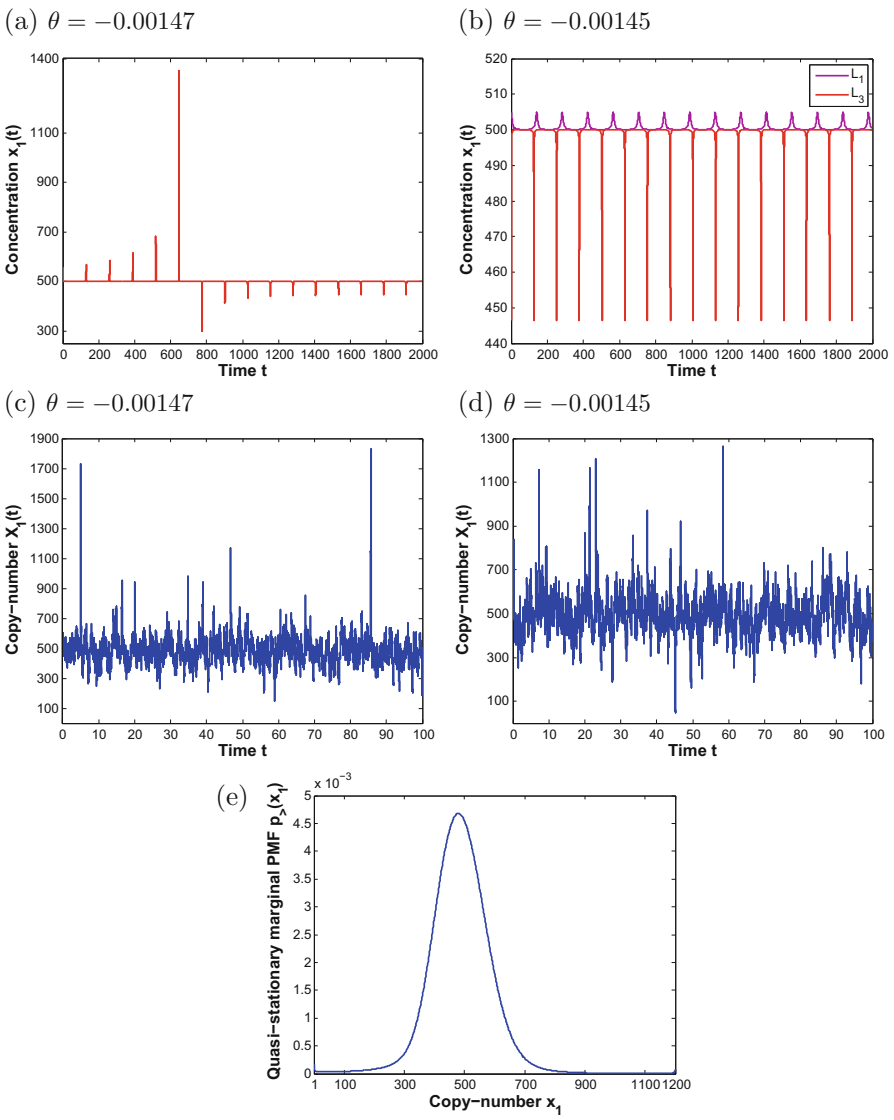


Fig. 4 (**a**)–(**b**) Numerical solutions of the kinetic ODE system given by (12) before and after the bifurcation, where in (**b**) the trajectory initiated near the stable limit cycle L_1 is shown in *purple*, while the one initiated near L_3 in *red*. (**c**)–(**d**) Sample paths generated by the Gillespie stochastic simulation algorithm applied to the induced reaction network (3) before and after the bifurcation. (**e**) Approximate quasi-stationary x_1-marginal PMF, obtained by numerically solving the stationary CME, corresponding to network (3), on the bounded domain $(x_1, x_2) \in [1, 1200] \times [1, 1200]$, for the same parameters values as in (**d**). The parameters appearing in (13) are fixed to $a = 1, b = -1$, $c = 0.5$, $d = 0.08$, $x_1^* = -3$, $\mathcal{T}_1 = \mathcal{T}_2 = 1000$, $\varepsilon = 0.01$, with the reactor volume $V = 0.5$, and θ as indicated in the plots. Coefficients (13) are multiplied by a constant factor of 2×10^{-5} (time-rescaling)

4 Summary

In the first part of the paper, in Sect. 2, we have presented theoretical results regarding oscillations, oscillation-related bifurcations and multistability in the planar quadratic kinetic ODEs (6), which are (appropriately) bounded in the nonnegative quadrant. Such ODEs are used in applications to describe the deterministic dynamics of concentrations of two biological/chemical species, with at most quadratic interactions. While the kinetic ODEs (6) inherit many properties from the more general planar quadratic ODEs (5), some properties, which are of biological/chemical relevance, are not necessarily inherited. For example, we have formulated the following open problem: while general planar quadratic ODEs (5) may display bicyclicity (a coexistence of two stable oscillatory attractors), is the same true for the kinetic planar quadratic ODEs (6)?

In Sect. 3, building upon the results from Sect. 2, and using the results from [2] and Appendix 1, we have constructed two reaction networks, with the deterministic dynamics described by planar cubic kinetic ODEs. The first network is given by (1), and, at the deterministic level, displays a homoclinic bifurcation, and a coexistence of a stable critical point and a stable limit cycle (mixed bistability). The second network is given by (3), and, at the deterministic level, displays a multiple limit cycle bifurcation, and a coexistence of two stable limit cycles (bicyclicity). The phase planes of the kinetic ODEs induced by the first network before and after the bifurcation are shown in Fig. 2a, b, respectively, while for the second network in Fig. 2c, d.

In Fig. 3, we have compared the deterministic and stochastic solutions corresponding to the first reaction network (1), with the rate coefficients such that the deterministic solutions are close to the homoclinic bifurcation. Analogously, in Fig. 4, we have done the same for reaction network (3), when the deterministic solutions are close to the multiple limit cycle bifurcation. In both Figs. 3 and 4, we observe qualitative differences between the deterministic and stochastic dynamics. In particular, the stochastic dynamics in Fig. 3 may display stochastic switching near the deterministic bifurcation. Furthermore, the dynamics of both networks are not affected qualitatively by the deterministic bifurcation sharply at the bifurcation point.

In Sect. 1, we have outlined the statistical inference problem, consisting of detecting and classifying cycles (oscillations) in noisy time-series, and we have put forward networks (1) and (3) as suitable test problems. Network (1) poses two inference challenges: firstly, let us consider the scenario shown in Fig. 3d–f. In this case, the relative separation between the two deterministic attractors is larger. Consequently, at the stochastic level, the corresponding marginal probability mass function (PMF), shown in Fig. 3f, is bimodal. However, the left peak, corresponding to the deterministic limit cycle, is much smaller than the right peak, corresponding to the deterministic critical point (a node). Using the shape of the marginal PMF, as put forward in [1], one cannot conclude the presence of a noisy limit cycle. Let us note that, by sufficiently increasing the distance between the two attractors,

the left PMF peak from Fig. 3f approximately vanishes, making the inference problem even harder. On the other hand, using the covariance function (and spectral analysis), as put forward in [1], may also be limited, as the noisy time-series spends a smaller amount of time near the deterministic limit cycle, as demonstrated in Fig. 3e. Secondly, let us consider the scenario shown in Fig. 3a–c, when the relative separation between the two deterministic attractors is smaller. In this case, it may be a challenge to infer that there are two distinct attractors 'hidden' in the time-series shown in Fig. 3b, and the PMF shown in Fig. 3c. The fact that the PMF in Fig. 3c is a non-Gaussian may be used as an indication of a certain dynamical complexity. The problem becomes more difficult for network (3), with two stable deterministic limit cycles 'hidden' in the noisy time-series shown in Fig. 4d, and in the PMF shown in Fig. 4e. Let us note that the PMF is approximately Gaussian, and this persists for a wide range of larger reactor volumes.

Acknowledgements The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme "Stochastic Dynamical Systems in Biology: Numerical Methods and Applications" where work on this paper was undertaken. This work was supported by EPSRC grant no EP/K032208/1. This work was partially supported by a grant from the Simons Foundation. Tomáš Vejchodský would like to acknowledge the institutional support RVO 67985840. Radek Erban would also like to thank the Royal Society for a University Research Fellowship.

Appendix 1: Perturbed *x*-Factorable Transformation

Definition 1 Consider applying an x-factorable transformation, as defined in [2], on (5), and then adding to the resulting right-hand side a zero-degree term $\varepsilon\mathbf{v}$, with $\varepsilon \geq 0$ and vector $\mathbf{v} = (1, 1)^\top$, resulting in

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \varepsilon\mathbf{v} + \mathcal{X}(\mathbf{x})\boldsymbol{\mathcal{P}}(\mathbf{x}; \mathbf{k}) = \varepsilon\mathbf{v} + (\Psi_{\mathcal{X}}\boldsymbol{\mathcal{P}})(\mathbf{x}; \mathbf{k}) \equiv (\Psi_{\mathcal{X}_\varepsilon}\boldsymbol{\mathcal{P}})(\mathbf{x}; \mathbf{k}). \tag{20}$$

Then $\Psi_{\mathcal{X}_\varepsilon} : \mathbb{P}_2(\mathbb{R}^2; \mathbb{R}^2) \to \mathbb{P}_3(\mathbb{R}^2; \mathbb{R}^2)$, mapping $\boldsymbol{\mathcal{P}}(\mathbf{x}; \mathbf{k})$ to $(\Psi_{\mathcal{X}_\varepsilon}\boldsymbol{\mathcal{P}})(\mathbf{x}; \mathbf{k})$, is called a *perturbed x-factorable transformation* if $\varepsilon \neq 0$. If $\varepsilon = 0$, the transformation reduces to an (unperturbed) x-factorable transformation, $\Psi_{\mathcal{X}} \equiv \Psi_{\mathcal{X}_0}$, defined in [2].

Lemma 1 $(\Psi_{\mathcal{X}_\varepsilon}\boldsymbol{\mathcal{P}})(\mathbf{x}; \mathbf{k})$ *from* Definition 1 *is a* kinetic function*, i.e.* $(\Psi_{\mathcal{X}_\varepsilon}\boldsymbol{\mathcal{P}})(\mathbf{x}; \mathbf{k}) \in \mathbb{P}_3^{\mathcal{K}}(\mathbb{R}^2_{\geq}; \mathbb{R}^2)$.

Proof $(\Psi_{\mathcal{X}}\boldsymbol{\mathcal{P}})(\mathbf{x}; \mathbf{k})$ is a kinetic function [2]. Since, from (20), $(\Psi_{\mathcal{X}_\varepsilon}\boldsymbol{\mathcal{P}})(\mathbf{x}; \mathbf{k}) = \varepsilon\mathbf{v} + (\Psi_{\mathcal{X}}\boldsymbol{\mathcal{P}})(\mathbf{x}; \mathbf{k})$, with $\varepsilon \geq 0$ and $\mathbf{v} = (1, 1)^\top$, it follows that $(\Psi_{\mathcal{X}_\varepsilon}\boldsymbol{\mathcal{P}})(\mathbf{x}; \mathbf{k})$ is kinetic as well. □

We now provide a theorem relating location, stability and type of the positive critical points of (5) and (20).

Theorem 1 *Consider the ODE system* (5) *with positive critical points* $\mathbf{x}^* \in \mathbb{R}^2_>$. *Let us assume that* $\mathbf{x}^* \in \mathbb{R}^2_>$ *is hyperbolic, and is not the degenerate case between a node and a focus, i.e. it satisfies the condition*

$$\left(\text{tr}\left(\nabla \boldsymbol{\mathcal{P}}(\mathbf{x}^*;\mathbf{k})\right)\right)^2 - 4\text{det}\left(\nabla \boldsymbol{\mathcal{P}}(\mathbf{x}^*;\mathbf{k})\right) \neq 0, \tag{21}$$

as well as conditions (ii) *and* (iii) *of* Theorem 3.3 *in* [2]. *Then positivity, stability and type of the critical point* $\mathbf{x}^* \in \mathbb{R}^2_>$ *are invariant under the perturbed x-factorable transformations* $\Psi_{\mathcal{X}_\varepsilon}$, *for sufficiently small* $\varepsilon \geq 0$. *Assume* (5) *does not have boundary critical points. Consider the two-dimensional ODE system* (20) *with* $\varepsilon = 0$, *and with boundary critical points denoted* $\bar{\mathbf{x}}^0 \in \mathbb{R}^2_\geq$, $\bar{\mathbf{x}}^0 = (\bar{x}^0_{b,1}, \bar{x}^0_{b,2})$, $\bar{x}^0_{b,1}\bar{x}^0_{b,2} = 0$. *Assume that for* $i \in \{1,2\}$

$$\frac{\partial \mathcal{P}_i(\bar{\mathbf{x}}^0_b;\mathbf{k})}{\partial x_i} \neq 0, \qquad \text{if } \bar{x}^0_{b,i} \neq 0, \tag{22}$$

and that for some $i \in \{1,2\}$

$$\mathcal{P}_i(\bar{\mathbf{x}}^0_b;\mathbf{k}) > 0, \qquad \text{if } \bar{x}^0_{b,i} = 0. \tag{23}$$

Then, the critical point $\bar{\mathbf{x}}^0_b \in \mathbb{R}^2_\geq$ *of the two-dimensional ODE system* (20) *with* $\varepsilon = 0$ *becomes the critical point* $\bar{\mathbf{x}}_b \notin \mathbb{R}^2_\geq$ *of system* (20) *for sufficiently small* $\varepsilon > 0$.

Proof The critical points of (20) are solutions of the following regularly perturbed algebraic equation

$$\varepsilon \mathbf{v} + \mathcal{X}(\bar{\mathbf{x}})\boldsymbol{\mathcal{P}}(\bar{\mathbf{x}};\mathbf{k}) = \mathbf{0}. \tag{24}$$

Let us assume $\bar{\mathbf{x}}$ can be written as the power series

$$\bar{\mathbf{x}} = \bar{\mathbf{x}}^0 + \varepsilon\bar{\mathbf{x}}^1 + \mathcal{O}(\varepsilon^2), \tag{25}$$

where $\bar{\mathbf{x}}^0 \in \mathbb{R}^2_\geq$ are the critical points of (20) with $\varepsilon = 0$. Substituting the power series (25) into (24), and using the Taylor series theorem on $\boldsymbol{\mathcal{P}}(\bar{\mathbf{x}};\mathbf{k})$, so that $\boldsymbol{\mathcal{P}}(\bar{\mathbf{x}}^0 + \varepsilon\bar{\mathbf{x}}^1 + \mathcal{O}(\varepsilon^2);\mathbf{k}) = \boldsymbol{\mathcal{P}}(\bar{\mathbf{x}}^0;\mathbf{k}) + \varepsilon\nabla\boldsymbol{\mathcal{P}}(\bar{\mathbf{x}}^0;\mathbf{k})\bar{\mathbf{x}}^1 + \mathcal{O}(\varepsilon^2)$, as well as that $\mathcal{X}(\bar{\mathbf{x}}) = \mathcal{X}(\bar{\mathbf{x}}^0) + \varepsilon\mathcal{X}(\bar{\mathbf{x}}^1) + \mathcal{O}(\varepsilon^2)$, and equating terms of equal powers in ε, the following system of polynomial equations is obtained:

$$\begin{aligned} &\mathcal{O}(1) : \mathcal{X}(\bar{\mathbf{x}}^0)\boldsymbol{\mathcal{P}}(\bar{\mathbf{x}}^0;\mathbf{k}) = 0, \\ &\mathcal{O}(\varepsilon) : \mathcal{X}(\bar{\mathbf{x}}^0)\nabla\boldsymbol{\mathcal{P}}(\bar{\mathbf{x}}^0;\mathbf{k})\bar{\mathbf{x}}^1 + \mathcal{X}(\bar{\mathbf{x}}^1)\boldsymbol{\mathcal{P}}(\bar{\mathbf{x}}^0;\mathbf{k}) = -\mathbf{v}. \end{aligned} \tag{26}$$

Order 1 *equation*. The positive critical points $\bar{\mathbf{x}}^0 \in \mathbb{R}^2_>$ satisfy $\boldsymbol{\mathcal{P}}(\bar{\mathbf{x}}^0;\mathbf{k}) = \mathbf{0}$. Since $\boldsymbol{\mathcal{P}}(\mathbf{x};\mathbf{k})$ has no boundary critical points by assumption, critical points $\bar{\mathbf{x}}^0_b \in \mathbb{R}^2_\geq$ with $\bar{x}^0_{b,i} = 0$, $\bar{x}^0_{b,j} \neq 0$, $\bar{x}^0_{b,1}\bar{x}^0_{b,2} = 0$, $i,j \in \{1,2\}$, satisfy $\mathcal{P}_i(\bar{\mathbf{x}}^0_b;\mathbf{k}) \neq 0$, $\mathcal{P}_j(\bar{\mathbf{x}}^0_b;\mathbf{k}) = 0$.

Order ε equation. Vector $\bar{\mathbf{x}}^1$, corresponding to a positive $\bar{\mathbf{x}}^0$, satisfies

$$\mathcal{X}(\bar{\mathbf{x}}^0)\nabla\boldsymbol{\mathcal{P}}(\bar{\mathbf{x}}^0;\,\mathbf{k})\bar{\mathbf{x}}^1 = -\mathbf{v},$$

which can be solved provided $\bar{\mathbf{x}}^0$ is a hyperbolic critical point. Vector $\bar{\mathbf{x}}_b^1$, corresponding to a nonnegative $\bar{\mathbf{x}}_b^0$, is given by

$$\bar{x}_{b,i}^1 = \begin{cases} -(\mathcal{P}_i(\bar{\mathbf{x}}_b^0;\,\mathbf{k}))^{-1}, & \text{if} \bar{x}_{b,i}^0 = 0, \\ \left(\frac{\partial\mathcal{P}_i(\bar{\mathbf{x}}_b^0;\,\mathbf{k})}{\partial x_i}\right)^{-1}\left((\mathcal{P}_j(\bar{\mathbf{x}}_b^0;\,\mathbf{k}))^{-1}\frac{\partial\mathcal{P}_i(\bar{\mathbf{x}}_b^0;\,\mathbf{k})}{\partial x_j} - (\bar{x}_{b,i}^0)^{-1}\right), & \text{if} \bar{x}_{b,i}^0 \neq 0, \end{cases}$$

from which conditions (22) and (23) follow. □

Appendix 2: Bicyclic System with Large Attractors

Consider the following deterministic kinetic equations

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= k_1 + x_1(-k_2 + k_3x_1 + k_4x_2 - k_5x_1x_2), \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= k_6 + x_2(k_7 - k_8x_1 + k_9x_2 + k_{10}x_1^2 - k_{11}x_2^2), \end{aligned} \tag{27}$$

with the coefficients $\mathbf{k}$ given by

$$\begin{aligned} &k_1 = 10^{-3}, \quad k_2 = 10, \quad k_3 = 1, \quad k_4 = 1, \quad k_5 = 0.1, \quad k_6 = 10^{-3}, \\ &k_7 = 3.7, \quad k_8 = 1.9, \quad k_9 = 1.01, \quad k_{10} = 0.1, \quad k_{11} = 0.05. \end{aligned} \tag{28}$$

The canonical reaction network induced by system (27), involving two species s_1 and s_2 and eleven reactions $r_1, r_2, \ldots, r_{11}$ under mass-action kinetics, is given by

$$\begin{aligned} r_1 &: \quad \varnothing \xrightarrow{k_1} s_1, & r_6 &: \quad \varnothing \xrightarrow{k_6} s_2, \\ r_2 &: \quad s_1 \xrightarrow{k_2} \varnothing, & r_7 &: \quad s_2 \xrightarrow{k_7} 2s_2, \\ r_3 &: \quad 2s_1 \xrightarrow{k_3} 3s_1, & r_8 &: \quad s_1 + s_2 \xrightarrow{k_8} s_1, \\ r_4 &: \quad s_1 + s_2 \xrightarrow{k_4} 2s_1 + s_2, & r_9 &: \quad 2s_2 \xrightarrow{k_9} 3s_2, \\ r_5 &: \quad 2s_1 + s_2 \xrightarrow{k_5} s_1 + s_2, & r_{10} &: \quad 2s_1 + s_2 \xrightarrow{k_{10}} 2s_1 + 2s_2, \\ & & r_{11} &: \quad 3s_2 \xrightarrow{k_{11}} 2s_2. \end{aligned} \tag{29}$$

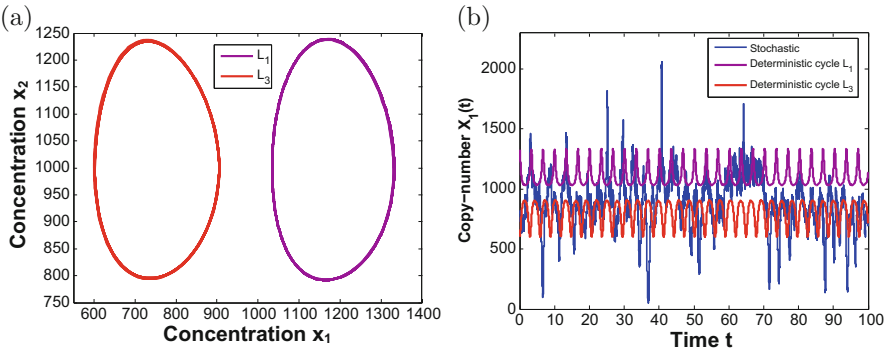


Fig. 5 Panel (**a**) displays numerically approximated stable limit cycles L_1 and L_3 in the state-space of system (27), with parameters (28) and reactor volume $V = 100$. Panel (**b**) displays in *blue* a representative sample path, generated by applying the Gillespie algorithm on the underlying reaction network (29) for the same parameters as in panel (**a**). Also shown are two deterministic trajectories, one initiated near the limit cycle L_1, while the other near L_3. One can observe that the stochastic sample path switches between the two deterministic attractors

In Fig. 5a, we show the two stable limit cycles obtained by numerically solving (27) with parameters (28). In Fig. 5b, in addition to the limit cycles, we also show in blue a representative sample path obtained by applying the Gillespie algorithm on (29). Let us note that (27) was constructed in a similar fashion as system (12) in Sect. 3.2, using the results from [40, 46].

References

1. M. Pineda-Krch, H.J. Blok, U. Dieckmann, M. Doebeli, A tale of two cycles – distinguishing quasi-cycles and limit cycles in finite predator-prey populations. Oikos **116**(1), 53–64 (2007)
2. T. Plesa, T. Vejchodský, R. Erban, Chemical reaction systems with a homoclinic bifurcation: an inverse problem. J. Math. Chem. **54**(10), 1884–1915 (2016)
3. P. Érdi, J. Tóth, *Mathematical Models of Chemical Reactions. Theory and Applications of Deterministic and Stochastic Models* (Manchester University Press/Princeton University Press, Princeton, 1989)
4. D.L.K. Toner, R. Grima, Molecular noise induces concentration oscillations in chemical systems with stable node steady states. J. Chem. Phys. **138**, 055101 (2013)
5. S. Louca, M. Doebeli, Distinguishing intrinsic limit cycles from forced oscillations in ecological time series. Theor. Ecol. **7**(4), 381–390 (2014)
6. R. Erban, S.J. Chapman, I. Kevrekidis, T. Vejchodský, Analysis of a stochastic chemical system close to a SNIPER bifurcation of its mean-field model. SIAM J. Appl. Math. **70**(3), 984–1016 (2009)
7. S. Liao, T. Vejchodský, R. Erban, Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks. J. R. Soc. Interface **12**(108), 20150233 (2015)
8. P. Thomas, A.V. Straube, J. Timmer, C. Fleck, R. Grima, Signatures of nonlinearity in single cell noise-induced oscillations. J. Theor. Biol. **335**, 222–234 (2013)

9. W. Vance, J. Ross, Fluctuations near limit cycles in chemical reaction systems. J. Chem. Phys. **105**, 479–487 (1996)
10. R.P. Boland, T. Galla, A.J. McKane, How limit cycles and quasi-cycles are related in systems with intrinsic noise. J. Stat. Mech. Theory Exp. **2008**, P09001, 1–27 (2008)
11. T. Xiao, J. Ma, Z. Hou, H. Xin, Effects of internal noise in mesoscopic chemical systems near Hopf bifurcation. New J. Phys. **9**, 403 (2007)
12. M.T. Borisuk, J.J. Tyson, Bifurcation analysis of a model of mitotic control in frog eggs. J. Theor. Biol. **195**, 69–85 (1998)
13. M.Y. Li, H. Shu, Multiple stable periodic oscillations in a mathematical model of CTL response to HTLV-I infection. Bull. Math. Biol. **73**, 1774–1793 (2011)
14. A. Amiranashvili, N.D. Schnellbächer, U.S. Schwarz, Stochastic switching between multi-stable oscillation patterns of the Min-system. New J. Phys. **18**, 093049 (2016)
15. F. Schlögl, Chemical reaction models for nonequilibrium phase transition. Z. Physik. **253**(2), 147–161 (1972)
16. V.A. Gaiko, *Global Bifurcation Theory and Hilbert's Sixteenth Problem* (Springer Science, New York, 2003)
17. V.A. Gaiko, On the geometry of polynomial dynamical systems. J. Math. Sci. **157**(3), 400–412 (2009)
18. L.M. Perko, Limit cycles of quadratic systems in the plane. Rocky Mt. J. Math. **14**(3), 619–645 (1984)
19. L.A. Cherkas, J.C. Artés, J. Llibre, Quadratic systems with limit cycles of normal size. *Buletinul Academiei de Ştiinţe a Republicii Moldova. Matematica* **1**(41), 31–46 (2003)
20. J.C. Artés, J. Llibre, Quadratic vector fields with a weak focus of third order. Publ. Math. **41**, 7–39 (1997)
21. C. Escher, Bifurcation and coexistence of several limit cycles in models of open two-variable quadratic mass-action systems. Chem. Phys. **63**, 337–348 (1981)
22. L.M. Perko, *Differential Equations and Dynamical Systems*, 3rd edn. (Springer, New York, 2001)
23. A.K. Dutt, Asymptotically stable limit cycles in a model of glycolytic oscillations. Chem. Phys. Lett. **208**, 139–142 (1992)
24. S. Kar, W.T. Baumann, M.R. Paul, J.J. Tyson, Exploring the roles of noise in the eukaryotic cell cycle. Proc. Natl. Acad. Sci. U. S. A. **106**, 6471–6476 (2009)
25. J.M.G. Vilar, H.Y. Kueh, N. Barkai, S. Leibler, Mechanisms of noise-resistance in genetic oscillators. Proc. Natl. Acad. Sci. U. S. A. **99**(9), 5988–5992 (2002)
26. Y.A. Kuznetsov, *Elements of Applied Bifurcation Theory*, 2nd edn. (Springer, New York, 2000)
27. M.S. Ghomi, A. Ciliberto, S. Kar, B. Novak, J.J. Tyson, Antagonism and bistability in protein interaction networks. J. Theor. Biol. **218**, 209–218 (2008)
28. Y. Dublanche, K. Michalodimitrakis, N. Kummerer, M. Foglierini, L. Serrano, Noise in transcription negative feedback loops: simulation and experimental analysis. Mol. Syst. Biol. **2**(41), E1–E12 (2006)
29. N. Bautin, On the number of limit cycles which appear with a variation of coefficients from an equilibrium position of focus or center type. Am. Math. Soc. Transl. **100**, 3–19 (1954)
30. M. Han, H. Zhu, The loop quantities and bifurcations of homoclinic loops. J. Diff. Equ. **234**, 339–359 (2007)
31. W. Coppel, A survey of quadratic systems. J. Diff. Equ. **2**, 293–304 (1966)
32. C. Chicone, T. Jinghuang, On general properties of quadratic systems. Am. Math. Mon. **89**, 167–178 (1982)
33. C. Escher, Double Hopf-bifurcation in plane quadratic mass-action systems. Chem. Phys. **67**, 239–244 (1982)
34. C. Escher, Models of chemical reaction systems with exactly evaluable limit cycle oscillations. Z. Phys. B **35**, 351–361 (1979)
35. G.M. Guidi, A. Goldbeter, Bistability without hysteresis in chemical reaction systems: a theoretical analysis of irreversible transitions between multiple steady states. J. Phys. Chem. **101**, 9367–9376 (1997)

36. G.M. Guidi, A. Goldbeter, Bistability without hysteresis in chemical reaction systems: the case of nonconnected branches of coexisting steady states. J. Phys. Chem. **102**, 7813–7820 (1998)
37. J. Tóth, Multistationarity is neither necessary nor sufficient to oscillations. J. Math. Chem. **25**, 393–397 (1998)
38. R.J. Dickson, L.M. Perko, Bounded quadratic systems in the plane. J. Diff. Equ. **7**, 251–273 (1970)
39. G.D.F. Duff, Limit cycles and rotated vector fields. Ann. Math. **67**, 15–31 (1953)
40. C.-C. Tung, Positions of limit cycles of the system $\mathrm{d}x/\mathrm{d}t = \sum a_{ik}x^i y^k$, $\mathrm{d}y/\mathrm{d}t = \sum b_{ik}x^i y^k$, $0 \leq i + k \leq 2$. Sci. Sin. **8**, 151–171 (1959)
41. M. Feinberg, *Lectures on Chemical Reaction Networks* (Delivered at the Mathematics Research Center, University of Wisconsin, Madison, 1979).
42. N.G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Burlington, 2007)
43. R. Erban, S.J. Chapman, P. Maini, *A Practical Guide to Stochastic Simulations of Reaction-Diffusion Processes*. Lecture Notes (2007). Available as http://arxiv.org/abs/0704.1908
44. T. Plesa, K. Zygalakis, D.F. Anderson, R. Erban, *Noise Control for DNA Computing* (2017, submitted). https://arxiv.org/abs/1705.09392
45. M. Vellela, H. Qian, A quasistationary analysis of a stochastic chemical reaction: Keizer's paradox. Bull. Math. Biol. **69**, 1727–1746 (2007)
46. L.M. Perko, Rotated vector fields. J. Diff. Equ. **103**, 127–145 (1993)

Importance Sampling for Metastable and Multiscale Dynamical Systems

K. Spiliopoulos

1 Introduction

In this paper, we discuss recent developments on importance sampling methods for metastable dynamics that may also have multiple scales. Development of accelerated Monte Carlo methods for metastable, multiple-scale processes is of great interest. Importance sampling is a variance reduction technique in Monte Carlo simulation, which is especially relevant when dealing with rare events. Since its introduction, importance sampling has been one of the most popular techniques for rare event simulation. There is a vast literature of papers investigating its applications from a broad set of sciences including engineering, chemistry, physics, biology, finance, insurance, e.g., [1, 10, 28, 31, 32, 36, 40, 46, 53, 54].

Consider a sequence $\{X^\epsilon\}_{\epsilon>0}$ of random elements and assume that we want to estimate the probability $0 < p^\epsilon = \mathbb{P}[X^\epsilon \notin \mathcal{D} \cup \partial\mathcal{D}] \ll 1$ for a given set $\mathcal{D}$, such that the event $\{X^\epsilon \notin \mathcal{D} \cup \partial\mathcal{D}\}$ is unlikely for small ϵ. If closed form formulas are not available, or numerical approximations are either too crude or unavailable, then one has to resort in simulation. It is well known that standard Monte Carlo simulation techniques (i.e., using the unbiased estimator $\hat{p}^\epsilon = \frac{1}{N}\sum_{j=1}^{N} 1_{X^{\epsilon,j}\notin\mathcal{D}\cup\partial\mathcal{D}}$) perform rather poorly in the rare-event regime. As it is known, see, for example, [1], in order to achieve relative error smaller than one using standard Monte Carlo, one needs an effective sample size $N \approx 1/p^\epsilon$. In other words, for a fixed computational cost, relative errors grow rapidly as the event becomes more rare. Thus standard Monte Carlo is infeasible for rare-event simulation.

K. Spiliopoulos (✉)
Department of Mathematics and Statistics, Boston University, Boston, MA 02215, USA
e-mail: kspiliop@math.bu.edu.

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_2

The goal of importance sampling is to simulate the system under an alternative probability distribution $\bar{\mathbb{P}}$ instead of the original probability $\mathbb{P}$. Let's say, for example, that we are interested in the estimation of

$$\mathbb{E}_y\left[e^{-\frac{1}{\epsilon}h(X_T^\epsilon)}\right] \text{or} \mathbb{P}_y\left[\tau^\epsilon_{\mathcal{D}\cup\partial\mathcal{D}} \leq T\right] \tag{1}$$

where $h : \mathbb{R}^d \mapsto \mathbb{R}$ is a positive function, $T > 0, \epsilon > 0$, $y \in D$ is the initial point, $\tau^\epsilon_{\mathcal{D}\cup\partial\mathcal{D}}$ is exit time from the set $\mathcal{D} \cup \partial\mathcal{D}$, X^ϵ is a stochastic process modeling the dynamics. Also, notice that the probability above can be considered (modulo the important technical point of lack of continuity) as a special case of $\mathbb{E}_y[e^{-\frac{1}{\epsilon}h(X_T^\epsilon)}]$, when h is, for example, chosen such that $h(x) = 0$ for $x \notin \mathcal{D} \cup \partial\mathcal{D}$ and $h(x) = +\infty$ for $x \in \mathcal{D} \cup \partial\mathcal{D}$.

When rare events dominate, then standard Monte Carlo methods perform poorly in the small noise limit. Then, to estimate $\mathbb{E}_y[e^{-\frac{1}{\epsilon}h(X_T^\epsilon)}]$, one generates iid samples $X^\epsilon_{(k)}$ from $\bar{\mathbb{P}}$ and uses the importance sampling estimator

$$\frac{1}{N}\sum_{k=1}^{N} e^{-\frac{1}{\epsilon}h(X^\epsilon_{(k)})}\frac{d\mathbb{P}}{d\bar{\mathbb{P}}}(X^\epsilon_{(k)}). \tag{2}$$

The key question is the design of $\bar{\mathbb{P}}$ such that the second moment $\bar{\mathbb{E}}_y[e^{-\frac{1}{\epsilon}h(X_T^\epsilon)}(d\mathbb{P}/d\bar{\mathbb{P}})(X^\epsilon_\cdot)]^2$ (and hence the variance) is minimized. $\bar{\mathbb{E}}$ is the expectation operator under $\bar{\mathbb{P}}$. The choice of the appropriate alternative measure $\bar{\mathbb{P}}$ is closely related to certain Hamilton-Jacobi-Bellman (HJB) equations.

The first issue that we address is the effect of rest points (and metastability in general) on importance sampling. It turns out that when dealing with metastability, even seemingly reasonable schemes that are also asymptotically optimal may perform poorly in practice. This includes also changes of measure that try to enforce the simulated trajectories to follow large deviations most likely paths. The reason for the degradation in performance is the role of prefactors. Prefactors can become very important when rest points are included in the domain of interest for the simulation. Large deviations based change of measures may not account for the prefactors, as they rely on logarithmic asymptotics. We elaborate on these issues and discuss potential ways on how the issue can be addressed.

The second issue that we address is the effect of multiple scales on the design of provably efficient importance sampling methods. It turns out that when the dynamical system has widely separated multiple scales, then one can use averaging and homogenization techniques. However, as we shall see, it is not sufficient to base the design of importance sampling on the effective homogenized dynamics. The local information needs to be taken into account. Mathematically this is done using the so-called cell problem, or macroscopic problem, in the theory of periodic and random homogenization.

The rest of the article is summarized as follows. In Sect. 2 we review the classical large deviations theory and the setup of importance sampling for small noise diffusions. In Sect. 3 we discuss the effects of rest points, i.e. of stable and

unstable equilibrium points, in the design of importance sampling. We argue why asymptotic optimality may actually not mean good practical performance and we also argue that following large deviations most likely optimal paths may lead to poor performance. In addition, we present constructions that lead to guaranteed good performance. We supplement the theoretical arguments by simulation studies. We refer the interested reader to [15, 23] for more details. In Sects. 4 and 5, we address the design of importance sampling schemes in the presence of multiple scales. We construct asymptotically optimal schemes in the presence of multiple scales. To be more precise, in Sect. 4 we consider overdamped Langevin dynamics in periodic multiscale environments and we review the related large deviations theory and importance sampling theory, presenting simulation studies. The interested reader can also consult [21, 22]. In Sect. 5 we review recent developments in large deviations and importance sampling for multiscale dynamics in random environments, see also [49, 50]. In Sect. 6 we describe how one can combine the results of Sect. 3 with those of Sects. 4 and 5 and also review future directions.

For the sake of concreteness and for exposition purposes we restrict the presentation of this article in the case of diffusions with gradient drift and constant diffusivity, which also implies reversible diffusion dynamics. However, we mention that almost all of the arguments can and have been generalized to the case with general state dependent drift and diffusion coefficient, especially those about the effect of multiple scales on importance sampling, see [14, 22, 23, 49, 50]. For results in the infinitely dimensional case, we refer the interested reader to [45].

2 Review of Large Deviations and Importance Sampling Theory for Diffusions

Let us briefly review the setup for small noise diffusions in $\mathbb{R}^d$ (e.g., [22, 51]) *without* the effect of multiple scales. Let W_t be a standard d-dimensional Wiener process and consider

$$dX_t^\epsilon = -\nabla V(X_t^\epsilon)dt + \sqrt{\epsilon}\Gamma dW_t, \quad X_{t_0}^\epsilon = y. \tag{3}$$

Large deviations principle for the process X_t^ϵ is well known (e.g., [27]). In particular, the action functional for the process $X_t^\epsilon, t_0 \le t \le T$, in $\mathcal{C}([t_0, T])$ as $\epsilon \downarrow 0$ has the form $\frac{1}{\epsilon} S_{t_0 T}(\phi)$, where

$$S_{t_0 T}(\phi) = \begin{cases} \frac{1}{2}\int_{t_0}^T (\dot{\phi}_s + \nabla V(\phi_s))^T \left[\Gamma\Gamma^T\right]^{-1} (\dot{\phi}_s + \nabla V(\phi_s))ds, & \text{if} \phi \in \mathcal{AC}([t_0, T]) \\ +\infty, & \text{otherwise.} \end{cases} \tag{4}$$

Here $\mathcal{C}([t_0, T])$, $\mathcal{AC}([t_0, T])$ are the sets of continuous and absolutely continuous functions on $[t_0, T]$ respectively. Then, under fairly general conditions,

$$\mathbb{E}_y\left[e^{-\frac{1}{\epsilon}h(X_T^\epsilon)}\right] \approx e^{-\frac{1}{\epsilon}\inf\left\{S_{t_0 T}(\phi)+h(\phi_T):\phi,\phi_{t_0}=y\right\}}, \text{as} \epsilon \downarrow 0.$$

A simple application of Jensen's inequality together with Varadhan's integral lemma (e.g., [13, 27, 52]) shows that an asymptotically optimal $\bar{\mathbb{P}}$ should satisfy

$$\lim_{\epsilon\to 0}\epsilon\ln\bar{\mathbb{E}}\left[e^{-\frac{1}{\epsilon}h(X_T^\epsilon)}d\mathbb{P}/d\bar{\mathbb{P}}\right]^2=-2G(t_0,y),$$
$$\text{with} G(t,x)=\inf_{\phi\in\mathcal{AC}([t,T]),\phi_t=x}\{S_{tT}(\phi)+h(\phi_T)\}$$

Turning to importance sampling, for $\bar{\mathbb{P}}$ that are absolutely continuous with respect to $\mathbb{P}$, Girsanov's formula implies

$$\frac{d\bar{\mathbb{P}}}{d\mathbb{P}}=e^{-\frac{1}{2\epsilon}\int_0^T|v_s|^2ds+\frac{1}{\sqrt{\epsilon}}\int_0^T v_s dW_s} \tag{5}$$

where v_t is a progressively measurable process (control) such that the right-hand side is a martingale (with respect to an appropriate filtration). Under $\bar{\mathbb{P}}$, X^ϵ satisfies

$$dX_t^\epsilon=\left[-\nabla V(X_t^\epsilon)+\Gamma v_t\right]dt+\sqrt{\epsilon}\Gamma d\bar{W}_t,\quad\text{with}\quad\bar{W}_t=W_t-\frac{1}{\sqrt{\epsilon}}\int_{t_0}^t v_\rho d\rho \tag{6}$$

So, the problem is restricted to choosing the control v_t optimally (i.e., such that the second moment is minimized) and then using the estimator based on iid samples generated from $\bar{\mathbb{P}}$ under (6). Under appropriate conditions, the zero-variance (i.e., the best) change of measure is based on the control v_t given by the formula $v_t=\bar{u}(t,X_t^\epsilon)$ where $\bar{v}(t,x)=-\Gamma^T\nabla G^\epsilon(t,x)$ where $G^\epsilon(t,x)$, with terminal condition $G^\epsilon(T,x)=h(x)$, is the solution to the PDE, of HJB type:

$$\partial_t G^\epsilon(t,x)-\nabla V(x)\cdot\nabla G^\epsilon(t,x)-\frac{1}{2}\left|\Gamma^T\nabla G^\epsilon(t,x)\right|^2+\frac{\epsilon}{2}\text{tr}\left[\Gamma\Gamma^T\nabla^2G^\epsilon(t,x)\right]=0. \tag{7}$$

Since (7) is not tractable, it is standard approach to go to the viscosity limit $\epsilon\downarrow 0$. Then $G(t,x)=\lim_{\epsilon\downarrow 0}G^\epsilon(t,x)$ is the viscosity solution to the HJB equation with Hamiltonian

$$H(x,p)=\langle-\nabla V(x),p\rangle-\frac{1}{2}\left\|\Gamma^Tp\right\|^2$$

i.e., to the equation

$$\partial_t G(t,x)-\nabla V(x)\cdot DG(t,x)-\frac{1}{2}\left|\Gamma^TDG(t,x)\right|^2=0,\quad G(T,x)=h(x). \tag{8}$$

Notice that by control arguments, e.g., see [25], we can also write

$$G(t,x)=\lim_{\epsilon\downarrow 0}G^\epsilon(t,x)=\inf_{\phi\in\mathcal{AC}([t,T]),\phi_t=x}\{S_{tT}(\phi)+h(\phi_T)\}.$$

In fact, more is true. A smooth function $\bar{U}(t,x) : [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}$ is called a *subsolution* to the HJB equation (8) with $\epsilon = 0$ if

$$\partial_t \bar{U}(t,x) - \nabla V(x) \cdot \nabla \bar{U}(t,x) - \frac{1}{2} \left| \Gamma^T \nabla \bar{U}(t,x) \right|^2 \geq 0, \quad \bar{U}(T,x) \leq h(x). \tag{9}$$

It turns out (Theorem 4.1 in [22]) that appropriate, *smooth* subsolutions are enough. If $\bar{U}(t,x) \in \mathcal{C}^{1,1}([t_0,T] \times \mathbb{R}^d)$ satisfies (9) and the *feedback* control to use in (6) is $v_t = -\Gamma^T \nabla \bar{U}(t, X_t^\epsilon)$, then

$$G(t_0,y) + \bar{U}(t_0,y) \leq \liminf_{\epsilon \to 0} -\epsilon \ln \bar{\mathbb{E}} \left[e^{-\frac{1}{\epsilon} h(X_T^\epsilon)} \frac{d\mathbb{P}}{d\bar{\mathbb{P}}} \right]^2 \leq 2G(t_0,y). \tag{10}$$

Therefore, asymptotic optimality is attained if $\bar{U}$ satisfies $\bar{U}(t_0,y) = G(t_0,y) = \lim_{\epsilon \downarrow 0} G^\epsilon(t_0,y)$ since then lower and upper bound agree. The design and analysis of importance sampling schemes based on the systematic connection with subsolutions to the appropriate HJB and Isaacs equations goes back to [16, 17]. See also [4–7] for the closely related concept of Lyapunov inequalities.

The importance sampling simulation scheme in order to estimate $\theta^\epsilon(t_0,y) \doteq \mathbb{E}_{t_0,y} \left[e^{-\frac{1}{\epsilon} h(X_T^\epsilon)} \right]$ goes as follows. Let $X^{\epsilon,v}$ be the solution to the SDE

$$dX_t^{\epsilon,v} = \left(-\nabla V(X_t^{\epsilon,v}) + \Gamma v_t \right) dt + \sqrt{\epsilon} \Gamma dW_t, \quad X_{t_0}^{\epsilon,u} = y. \tag{11}$$

1. Consider $v_t = \bar{u}(t, X_t^{\epsilon,v}) = -\Gamma^T \nabla_x \bar{U}(t, X_t^{\epsilon,v})$ with $\bar{U}$ an appropriate subsolution, i.e., it satisfies (9)
2. Consider the estimator

$$\hat{\theta}^\epsilon(y) \doteq \frac{1}{N} \sum_{j=1}^{N} \left[e^{-\frac{1}{\epsilon} h(X_T^{\epsilon,v}(j))} Z_j^v \right] \tag{12}$$

where

$$Z_j^v \doteq e^{-\frac{1}{2\epsilon} \int_0^T \left\| \bar{u}\left(t, X_t^{\epsilon,v}(j)\right) \right\|^2 dt - \frac{1}{\sqrt{\epsilon}} \int_0^T \left\langle \bar{u}\left(t, X_t^{\epsilon,v}(j)\right), dW_t(j) \right\rangle}$$

and $(W(j), X^{\epsilon,v}(j))$ is an independent sample generated from (11) with control $v_t = \bar{u}\left(t, X_t^{\epsilon,v}(j)\right)$.

We conclude this section, with the remark that a choice of the control v_t based on a subsolution as defined by (9) only guarantees logarithmic asymptotic optimality and does not say something about the *important* effect of pre-factors. As we will see in Sect. 3, this can imply degradation in the performance of the algorithm in problems with metastability. When dealing with metastability issues, things may be even more problematic if one is using the exact solution to the association HJB

equation, $G(t,x)$. While this may not be a problem for problems that do not involve rest points (i.e., does not involve stable or unstable equilibrium points) in the domain of interest, it does become problematic when dealing with metastability issues.

Remark 2.1 Obtaining accurately the solution $G(t,x)$ to the HJB equation (8), analytical or numerical, is challenging in high dimensions. However, even if this were possible, the solution by itself is not always suitable for importance sampling when one is interested in computing escape or transition probabilities. The issue is that in these cases, the solution is a viscosity solution with a discontinuous derivative at the rest point (stable or unstable equilibrium points) and with negative definite generalized second derivative there. Physically, the exact solution to the HJB equation attempts at each point in time and space to force the simulated trajectories to follow a most likely large deviations optimal path. However, by standard control arguments, see [25], the discontinuity of the spatial derivative at the rest point implies that multiple most likely optimal paths exist. As a consequence, the noise can cause trajectories to return to a neighborhood of the origin, thereby producing large likelihood ratios. In Sect. 3.2, we will see that this is a serious issue, leading to poor performance, even in dimension one where one can solve the HJB equation analytically. Importance sampling, when dealing with state dependent metastable dynamical systems, needs to be addressed from a global point of view and not local.

3 The Effect of Rest Points on Importance Sampling

As it is shown, mathematically and numerically, in [15, 23, 48], in dynamical systems that exhibit metastable behavior standard simulation methods do not readily apply. Asymptotic optimality is *necessary but not sufficient* for good performance due to the non-trivial effect of the pre-factors. The pre-factor computations in [23, 48] prove that there is non-trivial interaction of parameters such as the strength of the noise ϵ and the terminal time T. We remark here that this is in contrast to escape probabilities for other well-studied problems, such as stochastic networks, e.g., [4, 6, 17–20], because there the proximity of the rest point has little impact on either the asymptotic rate of decay or the pre-exponential term.

These interactive effects vanish in the logarithmic limit as the noise goes to zero, but they have a significant effect on the performance of the algorithms. The following question immediately presents itself:

- Is it sufficient to have schemes that are only asymptotically logarithmical optimal, in the sense that the second moment of the estimator satisfies (10)? What about pre-factors? Are they truly negligible in practice in the rare event regime?
- Can we construct a subsolution $\bar{U}(t,x)$ that not only satisfies (9) but it also takes care of the prefactor effects?

3.1 Effects in the Prelimit

Let us demonstrate the effect of prefactors on the behavior of estimators in the following classical simple setting. Let us assume that the diffusion coefficient $\Gamma = I$, and that $x = O$ is the global minimum for $V(x)$. In particular, let us assume that $DV(O) = 0$ and that $DV(x) \neq 0$ for every $x \neq O$. Define

$$\mathcal{D} = \left\{x \in \mathbb{R}^d : 0 \leq V(x) < L\right\}$$

and let $A_c = \left\{x \in \mathbb{R}^d : V(x) = c\right\}$. Then for an initial point y such that $0 \leq V(y) < L$, let us assume that we want to estimate

$$\theta^\epsilon(t, y) = \mathbb{P}_{t,y}\left\{X^\epsilon \text{hits} A_L \text{beforetime} T\right\}.$$

A classical quantity if interest in metastability theory is the quasipotential, see [27]. The quasipotential with respect to the equilibrium point O is defined as follows

$$W(O, x) = \{S_{0T}(\phi) : \phi \in \mathcal{C}([0, T]), \phi(O) = 0, \phi(T) = x, T \in (0, \infty)\}$$

Under the assumptions of this section, the quasipotential is computable in closed form [27]: $W(O, x) = 2V(x) \text{for} x \in \{y \in \mathcal{D} \cap \partial\mathcal{D} : V(y) \leq \inf_{z \in \partial\mathcal{D}} V(z)\}$.

Now, if we define $\tau^\epsilon = \inf\left\{t > 0 : X_t^\epsilon \notin \mathcal{D}\right\}$, then, as it is shown in [27] we have that $\lim_{\epsilon \downarrow 0} \epsilon \ln \mathbb{E}\tau^\epsilon = \inf_{z \in \partial\mathcal{D}} W(O, z)$. Thus, the quasipotential allows to approximate exit times in the logarithmic large deviations regime, [27]. Many quantities in the theory of metastability are defined via the quasipotential. The quasipotential characterizes the leading asymptotics of exit times and exit probabilities, approximates transition rates for reversible and irreversible systems, and allows to qualitatively describe transitions between stable attractors if the system has many of them; see also [11, 12, 24, 27, 38, 39] for more details. These conclusions hold for both gradient and non-gradient cases, but in the gradient case the quasipotential is computable in closed form.

Turning now to importance sampling, it is easy to verify that the quasipotential is a stationary subsolution to the associated HJB equation (9) with $\epsilon = 0$, by adding an appropriate constant C in order to justify the necessary boundary and terminal conditions. In particular, $\bar{U}_{QP}(x) = 2L - W(O, x)$ defines a subsolution for (9). It turns out, see [23], that the quasipotential yields a reasonable change of measure if rest points are not part of the domain of interest. However, this is no longer true if rest points are included in the domain of interest.

Let us denote $Q^\epsilon(0, y; \bar{u}) = \bar{\mathbb{E}}[e^{-\frac{1}{\epsilon}h(X_T^\epsilon)} d\mathbb{P}/d\bar{\mathbb{P}}]^2$ to be the second moment of the estimator constructed using the control $\bar{u}$. Based now on the arguments of [23] one can prove the following representation for the second moment of the estimator based on the change of measure induced by the control $\bar{u}(t, x) = -\nabla \bar{U}_{QP}(x)$

$$-\epsilon \log Q^\epsilon(0, y; \bar{u}) = \inf_{v \in \mathcal{A}} \mathbb{E}\left[\frac{1}{2}\int_0^{\hat{\tau}^\epsilon} \|v(s)\|^2 ds - \int_0^{\hat{\tau}^\epsilon} \left\|\bar{u}(\hat{X}_s^\epsilon)\right\|^2 ds + \infty 1_{\{\hat{\tau}^\epsilon > T\}}\right]. \tag{13}$$

where $\hat{X}_s^\epsilon$ is the unique solution to the SDE

$$d\hat{X}_s^\epsilon = -DV(\hat{X}_s^\epsilon)ds + \left[\sqrt{\epsilon}dW_s - [\bar{u}(\hat{X}_s^\epsilon) - v(s)]ds\right]$$

with initial condition $\hat{X}_0^\epsilon = y$ and $\hat{\tau}^\epsilon$ is the first time that $\hat{X}^\epsilon$ exits from $\mathcal{D}$.

It is important to note that (13) provides a non-asymptotic representation for the second moment of the estimator. By the arguments of [23], we can choose a particular admissible control $v(s)$ in (13) so that the following takes place. Let T be large and let $0 < K < T$ so that the time interval $[0, T]$ is split into $[0, T-K)$ and $[T-K, T]$. Set $v(s) = 0$ for $s \in [0, T-K)$. The resulting dynamics for $\hat{X}^\epsilon$ is stable for $s \in [T-K, T]$ and with high probability the process will stay around the point y for $s \in [0, T-K)$. In the time interval $[T-K, T]$, we set $v(s)$ so that escape happens prior to T. Then, it can be shown that there are positive constants $C_1, C_2 < \infty$, so that

$$Q^\epsilon(0, y; \bar{u}) \geq e^{-\frac{1}{\epsilon}C_1 + C_2(T-K)}.$$

This bound indicates that if T is large, one may need to go to considerably small values of ϵ in order to achieve the theoretical optimal asymptotic performance. We also remark that if T is large (see Chap. 4 of [27]), $G(0, y)$ and $\bar{U}(y)$ get closer in value. Thus, by (10) and for large enough T, the particular importance sampling scheme is asymptotically optimal.

Hence, we have just seen an example where an importance sampling estimator is almost asymptotically optimal, but it does not perform that well pre-asymptotically due to the effect of the possibly long time horizon T and its interplay with ϵ.

3.2 *The Problems Arising When Following Large Deviations Asymptotically Most Likely Paths and a Remedy to the Problem*

The connection of change of measures with HJB equations via large deviations is well situated for a systematic treatment of dynamic importance sampling schemes for state dependent processes like diffusions (3). For small noise diffusions the theoretical framework of subsolutions to HJB equations and their use for Monte Carlo methods can be found in [22]. It was a common belief for some time that if the underlying stochastic process has a large deviations principle and if the change of measure is consistent with the large deviations asymptotically most likely path leading to the rare event (an open-loop control), then the resulting importance sampling scheme would be optimal. However, such heuristics have been shown to be unreliable in general and simple examples have been constructed showing the failure of the corresponding importance schemes even in very simple settings [29, 30]. This

is due to the presence of "rogue-trajectories," i.e., unlikely trajectories, that are likely enough to increase likelihood ratios to the point that the performance is comparable to standard Monte Carlo. This is especially true for metastability problems (i.e., when transitions between fixed points occur at suitable (large) timescales) where multiple nearly optimal paths may exist.

Use of dynamic changes of measure, i.e. based on feedback controls (time and location dependent) becomes important, see [15, 23]. However, even changes of measures that are based on feedback controls, that are consistent with large deviations and lead to asymptotically optimal change of measures can also be problematic in practice. We demonstrate this below in Table 2. Namely, as it turns out, in the presence of rest points and metastability, the prefactors may affect negatively the behavior of estimators even if one is using asymptotically optimal changes of measure in the spirit of (10). Hence, it becomes important to use dynamic change of measures that are based on subsolutions but lead to good performance even pre-asymptotically.

To that end, *novel* explicit simulation schemes are then constructed in [15, 23] that perform provably well *both* asymptotically and non-asymptotically, even when the simulation time is long. These constructions are based on large deviations asymptotics [8, 9, 27], stochastic control arguments and asymptotic expansions [24, 25] and detailed asymptotic analysis of the subsolution to the associated HJB in the neighborhood of the rest point where the potential can be thought of as being approximately quadratic. Essentially, due to the fact that near the rest point, the potential can be thought of as being approximately quadratic, one can hope to solve or to approximate the solution to the associated variational problem there. Then one needs to patch this solution together with the quasipotential based subsolution (which is a good subsolution away from the rest point) in the right way. Then, the combined subsolution, see $\bar{U}^{\delta}(t,x)$ in (15), turns out to be a good approximation to the zero variance change of measure. Such schemes lead to importance sampling algorithms with provably good performance for all small $\epsilon > 0$ and without suffering from bad prefactor effects.

In order to illustrate the point, let us briefly demonstrate such a construction in the case of dimension one, see [23]. So, let us assume that $V(x) = \frac{\lambda}{2}x^2$ with $\lambda > 0$ and let us assume that we study the problem of crossing a level set, say L, of the potential function $V(x)$. Here, we can compute $G(t,x)$ in closed form and we get

$$G(t,x) = \inf_{\phi_t = x, V(\phi_T) = L} \left\{ \frac{1}{2} \int_t^T \left\| \dot{\phi}_s + \lambda \phi_s \right\|^2 ds \right\} = \inf_{\hat{x} \in V^{-1}(L)} \lambda \frac{\left(\hat{x} - x e^{\lambda(t-T)}\right)^2}{1 - e^{2\lambda(t-T)}}. \quad (14)$$

Notice that $G(t,x)$ is also a viscosity solution to the $\epsilon = 0$ HJB equation (8) when supplemented with the appropriate boundary conditions. Hence, based on (10) a change of measure based on $G(t,x)$, i.e., using the control $u(t,x) = -\partial_x G(t,x)$, is expected to yield an asymptotically efficient estimator. While this is true, we will see below that this is not sufficient to yield good performance. The fact that the function $G(t,x)$ is not continuously differentiable in the domain of interest implies

that multiple optimal paths exist, which is an intuitive reason for the degradation in performance that will be demonstrated below.

However, by appropriately mollifying $G(t,x)$ and combining it with the quasipotential subsolution (as constructed in Sect. 3.1), one can construct a global subsolution which performs provably well even pre-asymptotically. The point is that $G(t,x)$ provides a good change of measure while near the rest point, whereas the quasipotential induced subsolution $\bar{U}_{QP}(x) = 2L - W(O,x)$ provides a good change of measure away from the rest point. There are a few more issues to deal with though. The first one is that $G(t,x)$ is discontinuous near $t = T$. The second one is that we need to put them together in a smooth way that will define a global subsolution.

Since $G(t,x)$ is discontinuous at $t = T$, we introduce two mollification parameters t^* and M that will be appropriately chosen as functions of ϵ. Motivated by the fact that $G(t,x)$ is a good subsolution near the equilibrium point, we fix another parameter $\hat{L} \in (0,L]$. In the one-dimensional case, it is easy to solve the equation $V(x^*) = \hat{L}$ and in particular we get that $x^* = \pm\hat{x}$ where $\hat{x} = \sqrt{\frac{2\hat{L}}{\lambda}}$. As a matter of fact, instead of using $G(t,x)$ directly, we set

$$F^M(t,x;\hat{x}) = \lambda \frac{\left(\hat{x} - xe^{\lambda(t-T)}\right)^2}{\frac{1}{M} + 1 - e^{2\lambda(t-T)}}$$

In order now to pass smoothly between the $\bar{U}_{QP}(x)$ and $F^M(t,x;\hat{x})$ or $F^M(t,x;-\hat{x})$ without violating the subsolution property, we use the exponential mollification, see [17]

$$U^\delta(t,x) = -\delta \log\left(e^{-\frac{1}{\delta}\bar{U}_{QP}(x)} + e^{-\frac{1}{\delta}\left[F^M(t,x;\hat{x}) + \bar{U}_{QP}(\hat{x})\right]} + e^{-\frac{1}{\delta}\left[F^M(t,x;-\hat{x}) + \bar{U}_{QP}(-\hat{x})\right]}\right)$$

It is easy to see that as $\delta \downarrow 0$

$$\lim_{\delta \downarrow 0} U^\delta(t,x) = \min\{\bar{U}_{QP}(x), F^M(t,x;\hat{x}), F^M(t,x;-\hat{x})\}$$

Clearly, if we choose $\hat{L} = L$, then we get $\bar{U}_{QP}(\hat{x}) = 0$. Based on these constructions, a provably efficient importance sampling scheme is constructed in [23], based on the subsolution

$$\bar{U}^\delta(t,x) = \begin{cases} \bar{U}_{QP}(x), & t > T - t^* \\ U^\delta(t,x), & t \le T - t^* \end{cases} \tag{15}$$

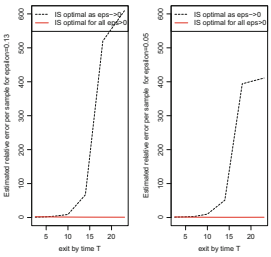
It turns out that $\bar{U}^\delta(t,x)$ is a global smooth subsolution which has provably good performance both pre-asymptotically and asymptotically. The role of the exponential mollification is to allow a smooth transition between the region that is near the equilibrium point and the region that is far away from it. The precise optimality bound and its proof guide the choice of the parameters δ, t^*, M, and $\hat{L}$.

Table 1 Parameter values for the algorithm based on a given value of $\epsilon > 0$

Parameter	δ	$\hat{L} \in (0, L]$	M	t^*
Values	2ϵ	$O(1)$ or ε^{2m} with $m < \kappa$	$\max\{\frac{\hat{L}}{\varepsilon^{2\kappa}}, 4\}$ with $\kappa \in (0, 1/2)$	$-\frac{2}{\lambda} \log \frac{1}{M}$

Table 2 *Left*: exit time distribution $\mathbb{P}_y \left[\tau^{\epsilon}_{\mathcal{D}\cup\partial\mathcal{D}} \leq T\right]$ for different pairs (ϵ, T), using the optimal change of measure constructed in [23]

$\epsilon \mid T$	2.5	7	10	18	23
0.20	$2e-02$	$8.3e-02$	$1.2e-01$	$2.1e-01$	$2.7e-01$
0.16	$7e-03$	$2.7e-02$	$4.0e-02$	$7.4e-02$	$9.5e-02$
0.13	$2e-03$	$6.9e-03$	$1.1e-02$	$2.0e-02$	$2.6e-02$
0.11	$4e-04$	$1.8e-03$	$2.8e-03$	$5.4e-03$	$7.0e-03$
0.09	$5e-05$	$2.6e-04$	$4.1e-04$	$7.8e-04$	$1.0e-03$
0.07	$2e-06$	$1.2e-05$	$1.9e-05$	$3.7e-05$	$4.8e-05$
0.05	$7e-09$	$4.4e-08$	$7.0e-08$	$1.4e-07$	$1.8e-07$



Events range from very rare to not so rare. Right: Comparison of relative errors per sample for two different changes of measure and for two values of ϵ. Small relative error is better

For the convenience of the reader, we present in Table 1 the suggested values for $(\delta, \hat{L}, M, t^*)$, given the value of the strength of the noise $\epsilon > 0$.

We refer the interested reader to [15, 23] for further details on the theoretical performance of the algorithm and on the choice of parameters.

In order to illustrate in a simple setting the effect of prefactors in the presence of metastable effects, we record in Table 2 Monte Carlo estimates based on $K = 10^7$ trajectories for the exit time distribution $\mathbb{P}_y \left[\tau^{\epsilon}_{\mathcal{D}\cup\partial\mathcal{D}} \leq T\right]$ from the basin of attraction of the left attractor of the potential of Fig. 1 for the process X^{ϵ} given by (3) with $\Gamma = I$. We used the importance sampling (IS) methods of [23], i.e., the change of measure based on the subsolution (15) and record estimates for different pairs (ϵ, T). In the figures next to Table 2, we compare the relative errors per sample of (a): the algorithm, which is optimal for all $\epsilon > 0$, i.e. the one based on the subsolution $\bar{U}^{\delta}(t, x)$, with (b): the IS algorithm that is consistent with the large deviations asymptotically most likely path leading to the rare event, i.e the one based on the actual solution $G(t, x)$ of the associated HJB equation. Notice however that the IS algorithm based on $G(t, x)$ is only asymptotically optimal in the large deviations logarithmic sense as $\epsilon \downarrow 0$ [i.e., it satisfies (10)].

Using relative error per sample as comparison criterium, we compare the two algorithms for two values of ϵ, one for which the events are not so rare ($\epsilon = 0.13$) and one for which the events are very rare ($\epsilon = 0.05$). Exact values are in the table, and we remark for completeness that intermediate behavior is qualitatively the same. Both algorithms perform well when T is small, but the algorithm that is based on the solution of the associated HJB equation, which is only logarithmic asymptotically optimal, starts deteriorating considerably as T gets large. The latter is an effect of the pre-factors becoming important. On the other hand, the change of measure constructed in [23] that takes into account the pre-factor information and is pre-asymptotically optimal yields optimal performance independently of the values

ϵ and T with relative errors around one, meaning that the values recorded at the table are reliable. It is important to note that due to large deviations, exit happens in long time scales, which implies that reliable estimates, especially when T is large, are essential.

4 Importance Sampling for Rough Energy Landscapes

In Sect. 3, we reviewed some of the practical issues that come up when one is trying to apply importance sampling techniques to metastable dynamics. While in Sect. 3 we ignored the effect of multiple scales, the goal of this section is to address the role of multiple scales in the design of asymptotically optimal importance sampling schemes.

A particular model of interest in chemical physics is the first order Langevin equation (16). Let us consider

$$dX_t^{\epsilon,\delta} = \left[-\frac{\epsilon}{\delta}\nabla Q\left(X_t^{\epsilon,\delta}/\delta\right) - \nabla V\left(X_t^{\epsilon,\delta}\right)\right]dt + \sqrt{\epsilon}\sqrt{2D}dW_t, \quad X_0^{\epsilon,\delta} = y, \quad ,0 < \epsilon, \delta \ll 1, \tag{16}$$

where the two-scale potential is composed by a large-scale part, $V(x)$, and a fluctuating part, $\epsilon Q(x/\delta)$. If Q is periodic, then we have a periodic environment, whereas if Q is random then we have a random environment. Models like (16) can be used to model rough energy landscapes [2, 21, 33, 55]. As it has been suggested (e.g., [37, 55]), the associated energy landscapes of certain biomolecules can be rugged (i.e., consist of many local "small" minima within local deep minima separated by barriers of varying heights). When one is interested in rare events, large deviations and Monte Carlo methods are relevant.

If $Q(y)$ is *periodic*, large deviations for multiscale diffusions in periodic environments are obtained in [14, 26, 47] for all possible interactions between ϵ and

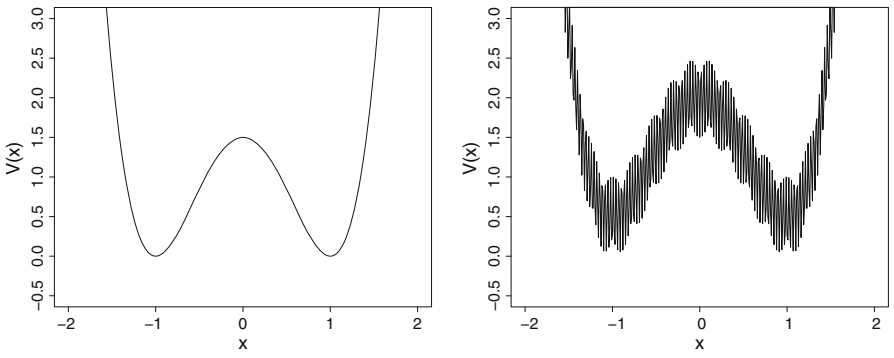


Fig. 1 A smooth and a rough potential function (energy landscape) with two wells

δ, setting the ground for the mathematical formulation of the related importance sampling theory, [21, 22, 47]. The *novel feature* is that the optimal change of measure for importance sampling is not based only on the gradient of the homogenized HJB equation (as in Sect. 2). The effect of fluctuations, which is quantified via the solution to the "cell problem" in homogenization [3, 43], is equally important. The cell problem is the solution to a Poisson type PDE. It is used to define the so called "corrector," which characterizes the first order correction in the approximation of the multiscale HJB by its homogenized limit. Therefore, when compared to the case without multiple scales, one needs more detailed information in order to guarantee, *at least*, asymptotic optimality.

For example, consider model (16) in the case $\frac{\epsilon}{\delta} \uparrow \infty$. Define the Gibbs measure

$$\mu(dy) = \frac{1}{L} e^{-\frac{Q(y)}{D}} dy, \quad L = \int_{\mathbb{T}^d} e^{-\frac{Q(y)}{D}} dy.$$

Then denote by $\chi(y)$ the smooth solution to the "cell problem"

$$-\nabla Q(y) \cdot \nabla \chi(y) + D\mathrm{tr}\left[\nabla^2 \chi(y)\right] = \nabla Q(y), \qquad \int \chi(y)\mu(dy) = 0. \tag{17}$$

The following large deviations result holds which is a special case of the results of [14]. In particular, [14] covers the case of general state dependent drift (not necessarily of gradient form) and state dependent diffusion coefficient.

Theorem 4.1 (Theorem 5.3 of [14] for the Case of (16)) *Assume that the functions $\nabla Q(y)$ and $\nabla V(x)$ are continuous and globally bounded, as are their partial derivatives up to order 1 in y and order 2 in x, respectively. Let $\{X^{\epsilon,\delta}, \epsilon, \delta > 0\}$ be the unique strong solution to (16). Let*

$$r(x) = -\int_{\mathbb{T}^d} \left(I + \frac{\partial \chi(y)}{\partial y}\right) \mu(dy) \nabla V(x),$$

$$q = 2D \int_{\mathbb{T}^d} \left(I + \frac{\partial \chi(y)}{\partial y}\right) \left(I + \frac{\partial \chi(y)}{\partial y}\right)^T \mu(dy),$$

where I denotes the identity matrix. If $\epsilon/\delta \to \infty$, then $\{X^{\epsilon,\delta}, \epsilon, \delta > 0\}$ converges in probability as $\epsilon, \delta \to 0$ to the solution of the ODE

$$d\bar{X}_t = r(\bar{X}_t)dt$$

and satisfies a large deviations principle with rate function

$$S_{tT}(\phi) = \begin{cases} \frac{1}{2} \int_t^T \left(\dot{\phi}_s - r(\phi_s)\right) q^{-1} \left(\dot{\phi}_s - r(\phi_s)\right)^T ds & \text{if} \phi \in \mathcal{AC}([t,T]), \phi_t = x \\ +\infty & \text{otherwise.} \end{cases}$$

In addition, it turns out that an asymptotically efficient change of simulation measure can be constructed analogously to Sect. 3, but based on the feedback control (see Theorem 4.1 in [22])

$$v_t = \bar{u}(t, X_t^\epsilon, X_t^\epsilon/\delta), \quad \text{with} \quad \bar{u}(t,x,y) = -\sqrt{2D}\,(I + \partial\chi(y)/\partial y)^T \nabla_x \bar{U}(t,x). \quad (18)$$

$\bar{U}(t,x)$ satisfies the inequalities in (9) with the homogenized (averaged) coefficients $r(x)$ and q in place of the original ones $-\nabla V(x)$ and $\Gamma = \sqrt{2D}I$ (compare with (9)). In particular, the second moment of an estimator with change of measure based on the control v_t by (18) will satisfy (10); this is Theorem 4.1 in [22].

Thus, compared to the case without multiscale features, one needs to consider the extra factor $(I + \partial\chi(y)/\partial y)$, that can be thought as the appropriate *weight* function, to achieve asymptotic optimality. In the absence of multiple scales, i.e., when $Q = 0$, we have $\chi = 0$ and we recover the case studied in Sect. 3. The numerical simulation studies of [21, 22] verify the need for accounting for the local environment via the weights $(I + \partial\chi(y)/\partial y)$ in the change of simulation measure.

Before illustrating the performance of this importance sampling scheme in a simulation study, let us demonstrate theoretically the necessity to include the cell problem information in the design of the change of measure. For simplicity purposes, let us restrict attention to dimension one. As we have seen before, the effective diffusion coefficient is given by

$$q = 2D \int_{\mathbb{T}} \left(1 + \frac{\partial\chi}{\partial y}\right)^2 \mu(dy)$$

In this case, the optimal change of measure is based on the control

$$\bar{u}(t,x,y) = -\sqrt{2D}\,(1 + \partial\chi(y)/\partial y)\,\partial_x \bar{U}(t,x).$$

So, let us assume that one is using instead the change of measure, based on the control dictated by the averaged dynamics. Namely, let us assume that the control in question is $\hat{u}(t,x) = -\sqrt{q}\partial_x \bar{U}(t,x)$.

A verification theorem, see [22] for details, shows that one would need a statement of the form

$$\text{“}\mathbb{E}\int_t^T \left[\sqrt{2D}\left(1 + \frac{\partial\chi}{\partial y}\left(\frac{X_s^{\epsilon,\delta}}{\delta}\right)\right) - \sqrt{q}\right] ds \to 0\text{”} \quad (19)$$

By averaging principle, this is true if

$$\sqrt{q} = \int_{\mathbb{T}} \sqrt{2D}\left(1 + \frac{\partial\chi(y)}{\partial y}\right)\mu(dy). \quad (20)$$

However, this is impossible, since

$$\left(\int \left(1 + \frac{\partial\chi(y)}{\partial y}\right)\mu(dy)\right)^2 \neq \int \left(1 + \frac{\partial\chi(y)}{\partial y}\right)^2 \mu(dy).$$

This last property explains mathematically why, the local information, as quantified via the cell problem, needs to be taken into account in the design of importance sampling. In Sect. 4.1, we will also see numerical evidence of this issue.

4.1 A Simulation Study

Let us demonstrate the performance of the importance sampling scheme in a simple simulation study. Consider the one well potential function with diffusion coefficient $D = 1$,

$$V(x) = \frac{1}{2}x^2, \qquad Q(y) = \cos(y) + \sin(y) \tag{21}$$

Assume that we want to estimate $\theta(\epsilon, \delta) = \mathbb{E}\left[e^{-\frac{1}{\epsilon}h(X_1^{\epsilon,\delta})}\right]$, where $h(x) = (|x| - 1)^2$.

It is easy to see that we are dealing with a rare event here, as the function $h(x)$ is minimized at $|x| = 1$. Let us compare the following three different estimators

$$\hat{\theta}_0(\epsilon, \delta) = \frac{1}{K}\sum_{j=1}^{K}\left[e^{-\frac{1}{\epsilon}h(X_1^{\epsilon,\delta}(j))}\right] - - -\text{standardMonteCarlo}$$

$$\hat{\theta}_1(\epsilon, \delta) = \frac{1}{K}\sum_{j=1}^{K}\left[e^{-\frac{1}{\epsilon}h(\bar{X}_1^{\epsilon,\delta,\bar{u}}(j))}Z_j^{\bar{u}}\right] - - -\text{optimal}$$

$$\hat{\theta}_2(\epsilon, \delta) = \frac{1}{K}\sum_{j=1}^{K}\left[e^{-\frac{1}{\epsilon}h(\bar{X}_1^{\epsilon,\delta,\hat{u}}(j))}Z_j^{\hat{u}}\right] - - -\text{ignoreslocalinformation}$$

where we have defined the controls

- $\bar{u}(t, x, y) = -\sqrt{2}\,(1 + \partial\chi(y)/\partial y)\,G_x(t, x)$—asymptotically optimal.
- $\hat{u}(t, x) = -\sqrt{q}G_x(t, x)$—based only on the homogenized system.

and the likelihood ratio is $Z_j^u = \frac{dP}{d\bar{P}}(\bar{X}_1^{\epsilon,\delta,u}(j))$. Notice that in this case, we can compute

$$1 + \frac{\partial\chi(y)}{\partial y} = e^{Q(y)} / \int_{\mathbb{T}} e^{Q(y)}dy,$$

which justifies the interpretation of the term $1 + \frac{\partial\chi(y)}{\partial y}$ as the proper weight term needed that takes into account the local information.

In Table 3, we see simulation studies based on $N = 10^7$ simulation trajectories each, for the estimation of $\theta(\epsilon, \delta)$ using the three different estimators. The measure of comparison is chosen to be the relative error per sample, defined to be

$$\hat{\rho}_i(\epsilon, \delta) \doteq \sqrt{N}\frac{\sqrt{\text{Var}(\hat{\theta}_i(\epsilon, \delta))}}{\hat{\theta}_1(\epsilon, \delta)}.$$

Table 3 Comparing different importance sampling estimators

No.	ϵ	δ	ϵ/δ	$\hat{\theta}_1(\epsilon,\delta)$	$\hat{\rho}_0(\epsilon,\delta)$	$\hat{\rho}_1(\epsilon,\delta)$	$\hat{\rho}_2(\epsilon,\delta)$
1	0.25	0.1	2.5	$2.25e-01$	1	6	20
2	0.125	0.04	3.125	$3.65e-02$	3	6	5
3	0.0625	0.015625	4	$8.75e-04$	34	4	13
4	0.03125	0.007	4.46	$6.87e-07$	141	3	105
5	0.025	0.004	6.25	$1.61e-08$	217	2	97
6	0.02	0.002	10	$1.99e-10$	1294	1	157
7	0.015	0.0013	11.54	$1.37e-13$	800	1	588

It is clear that the importance sampling scheme based on the asymptotically optimal change of measure $\bar{u}(t,x,y)$ outperforms the standard Monte Carlo estimator in which no change of measure is being done. It also outperforms, the estimator based solely on the homogenized system, which ignores the local information characterized by solution to the cell problem $\chi(y)$.

5 Importance Sampling for Multiscale Diffusions in Random Environments

Let $0 < \epsilon, \delta \ll 1$ and consider the process $(X^\epsilon, Y^\epsilon) = \left\{\left(X_t^\epsilon, Y_t^\epsilon\right), t \in [0,T]\right\}$ taking values in the space $\mathbb{R}^m \times \mathbb{R}^{d-m}$ that satisfies the system of SDEs

$$\begin{aligned} dX_t^\epsilon &= \left[\frac{\epsilon}{\delta} b\left(Y_t^\epsilon, \gamma\right) + c\left(X_t^\epsilon, Y_t^\epsilon, \gamma\right)\right] dt + \sqrt{\epsilon}\sigma\left(X_t^\epsilon, Y_t^\epsilon, \gamma\right) dW_t, \\ dY_t^\epsilon &= \frac{1}{\delta}\left[\frac{\epsilon}{\delta} f\left(Y_t^\epsilon, \gamma\right) + g\left(X_t^\epsilon, Y_t^\epsilon, \gamma\right)\right] dt + \frac{\sqrt{\epsilon}}{\delta}\left[\tau_1\left(Y_t^\epsilon, \gamma\right) dW_t + \tau_2\left(Y_t^\epsilon, \gamma\right) dB_t\right], \\ X_0^\epsilon &= x_0, \quad Y_0^\epsilon = y_0 \end{aligned} \tag{22}$$

We assume non-degeneracy of the diffusion coefficients as well $\mathcal{C}^1$ smoothness and boundedness of the drift and diffusion coefficients. Moreover, we assume that $\delta = \delta(\epsilon) \downarrow 0$ such that $\epsilon/\delta \uparrow \infty$ as $\epsilon \downarrow 0$. (W_t, B_t) is a 2κ-dimensional standard Wiener process. We assume that for each fixed $x \in \mathbb{R}^m$, $b(\cdot,\gamma), c(x,\cdot,\gamma), \sigma(x,\cdot,\gamma), f(\cdot,\gamma)$, $g(x,\cdot,\gamma), \tau_1(\cdot,\gamma)$ and $\tau_2(\cdot,\gamma)$ are stationary and ergodic random fields in an appropriate probability space $(\Gamma, \mathcal{G}, \nu)$ with $\gamma \in \Gamma$.

Example 5.1 Notice that if we choose $b(y,\gamma) = f(y,\gamma) = -\nabla_y Q(y,\gamma)$ for a periodic function $Q(\cdot)$, $c(x,y,\gamma) = -\nabla_x V(x)$, $\sigma(x,y,\gamma) = \tau_1(y,\gamma) = \sqrt{2D}$ and $\tau_2(y,\gamma) = 0$, and set $y_0 = x_0/\delta$, we then get the Langevin equation (16). In particular, if we make these choices, then we simply have $Y_t^\epsilon = X_t^\epsilon/\delta$ and the model

can be interpreted as diffusion in the rough potential $\epsilon Q(x/\delta, \gamma) + V(x)$, where the roughness is dictated by Q. In general, Q may not be modelled as a periodic function. One may model Q as a random field; see the simulation study in Sect. 5.3.

5.1 Description of the Random Environment

The large deviations and importance sampling results for (22), see [49, 50], are true under certain assumptions on the random medium that we recall here for convenience. We assume that there is a group of measure preserving transformations $\{\tau_y, y \in \mathbb{R}^{d-m}\}$ acting ergodically on Γ that is defined as follows.

Definition 5.2

i. τ_y preserves the measure, namely $\forall y \in \mathbb{R}^{d-m}$ and $\forall A \in \mathcal{G}$ we have $\nu(\tau_y A) = \nu(A)$.
ii. The action of $\{\tau_y : y \in \mathbb{R}^{d-m}\}$ is ergodic, that is if $A = \tau_y A$ for every $y \in \mathbb{R}^d$ then $\nu(A) = 0$ or 1.
iii. For every measurable function f on $(\Gamma, \mathcal{G}, \nu)$, the function $(y, \gamma) \mapsto f(\tau_y \gamma)$ is measurable on $\left(\mathbb{R}^{d-m} \times \Gamma, \mathbb{B}(\mathbb{R}^{d-m}) \otimes \mathcal{G}\right)$.

Let $\tilde{\phi}$ be a square integrable function in Γ and define the operator $T_y \tilde{\phi}(\gamma) = \tilde{\phi}(\tau_y \gamma)$. The operator $T_y \cdot$ is a strongly continuous group of unitary maps in $L^2(\Gamma)$, see [41]. Denote by D_i the infinitesimal generator of T_y in the direction i, which is a closed and densely defined generator, see [41].

In order to guarantee that the involved functions are ergodic and stationary random fields on $\mathbb{R}^{d-m}$, for $\tilde{\phi} \in L^2(\Gamma)$, let us define the operator $\phi(y, \gamma) = \tilde{\phi}(\tau_y \gamma)$. Similarly, for a measurable function $\tilde{\phi} : \mathbb{R}^m \times \Gamma \mapsto \mathbb{R}^m$ we consider the (locally) stationary random field $(x, y) \mapsto \tilde{\phi}(x, \tau_y \gamma) = \phi(x, y, \gamma)$. Then, it is guaranteed that $\phi(y, \gamma)$ (respectively, $\phi(x, y, \gamma)$) is a stationary (respectively, locally stationary) ergodic random field.

The coefficients, $b, c, \sigma, f, g, \tau_1, \tau_2$ of (22) are defined through this procedure and therefore are guaranteed to be ergodic and stationary random fields. For example, in the case of the c drift term, we start with an $L^2(\Gamma)$ function $\tilde{c}(x, \gamma)$ and we define the corresponding coefficients via the relation $c(x, y, \gamma) = \tilde{c}(x, \tau_y \gamma)$.

For every $\gamma \in \Gamma$, let us the operator

$$\mathcal{L}^\gamma = f(y, \gamma) \nabla_y \cdot + \text{tr}\left[\left(\tau_1(y, \gamma)\tau_1^T(y, \gamma) + \tau_2(y, \gamma)\tau_2^T(y, \gamma)\right) \nabla_y \nabla_y \cdot\right]$$

which is the infinitesimal generator of a Markov process, say $Y_{t,\gamma}$. Using the Markov process $Y_{t,\gamma}$, we can define the so-called environment process, see [35, 41, 42, 44], denoted by γ_t. The environment process γ_t has continuous transition probability densities with respect to the d-dimensional Lebesgue measure, see [41], and is defined by the equations

$$\gamma_t = \tau_{Y_{t,\gamma}}\gamma$$

$$\gamma_0 = \tau_{y_0}\gamma$$

The infinitesimal generator of the Markov process γ_t is given by

$$\tilde{L} = \tilde{f}(\gamma)D\cdot + \text{tr}\left[\left(\tilde{\tau}_1(\gamma)\tilde{\tau}_1^T(\gamma) + \tilde{\tau}_2(\gamma)\tilde{\tau}_2^T(\gamma)\right)D^2\cdot\right].$$

In order to simplify the presentation, let us assume that the operator $\tilde{L}$ is in divergence form. In particular, let us set $\tilde{f}(\gamma) = -D\tilde{Q}(\gamma)$ and $\tilde{\tau}_1(\gamma) = \sqrt{2D}\theta = $ constant and $\tilde{\tau}_2(\gamma) = \sqrt{2D}\sqrt{1-\theta^2} = $ constant.

Then, we can write the unique ergodic invariant measure for the environment process $\{\gamma_t\}_{t\geq 0}$ in closed form; see [41, 50] for more general case which is not necessarily restricted to the gradient case. Denote by $\mathbb{E}^\nu$ the expectation operator with respect to the measure ν. Then , the measure $\pi(d\gamma)$ defined on $(\Gamma, \mathcal{G})$ by

$$\pi(d\gamma) \doteq \frac{\tilde{m}(\gamma)}{\mathbb{E}^\nu \tilde{m}(\cdot)}\nu(d\gamma), \text{with} \tilde{m}(\gamma) = \exp[-\tilde{Q}(\gamma)/D].$$

is the unique ergodic invariant measure for the environment process $\{\gamma_t\}_{t\geq 0}$.

Next, we need to define the equivalent to the cell problem in the case of periodic coefficients, also known as the macroscopic problem in the homogenization theory. To do so, we first define $\mathcal{H}^1 = \mathcal{H}^1(\nu)$ to be the Hilbert space equipped with the inner product

$$(\tilde{f}, \tilde{g})_1 = \sum_{i=1}^{d}(D_i\tilde{f}, D_i\tilde{g}).$$

Let us consider $\rho > 0$ and consider the following problem on Γ

$$\rho\tilde{\chi}_\rho - \tilde{L}\tilde{\chi}_\rho = \tilde{b}. \tag{23}$$

Under the condition $\tilde{b} \in L^2(\nu)$ with $\left\|\tilde{b}\right\|_{\mathcal{H}^{-1}} < \infty$, Lax-Milgram lemma, see [34, 41], guarantees that Eq. (23) has a unique weak solution in the abstract Sobolev space $\mathcal{H}^1$ or equivalently in $\mathcal{H}^1(\pi)$. At this point, we note that in the periodic case one also considers (23), but one can then take $\rho = 0$ given that b averages to zero when is integrated against the invariant measure π. However, in the random case, (23) with $\rho = 0$ does not necessarily have a well-defined solution (even if b averages to zero when is integrated against the invariant measure π), see, for example, [34].

In the general random case, we consider the equation with $\rho > 0$ and in the end, the homogenization theorem is proven by taking appropriate sequences $\rho = \rho(\epsilon)$ such that $\rho(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$. Taking $\rho \downarrow 0$ is allowed by the following well-known properties of the solution to (23), (see [41, 42, 44]),

1. There is a constant K that is independent of ρ such that

$$\rho\mathbb{E}^{\pi}\left[\tilde{\chi}_{\rho}(\cdot)\right]^2 + \mathbb{E}^{\pi}\left[D\tilde{\chi}_{\rho}(\cdot)\right]^2 \leq K$$

2. $\tilde{\chi}_{\rho}$ has an $\mathcal{H}^1$ strong limit, i.e., there exists a $\tilde{\chi}_0 \in \mathcal{H}^1(\pi)$ such that

$$\lim_{\rho\downarrow 0}\left\|\tilde{\chi}_{\rho}(\cdot) - \tilde{\chi}_0(\cdot)\right\|_1 = 0 \quad \text{and} \quad \lim_{\rho\downarrow 0}\rho\mathbb{E}^{\pi}\left[\tilde{\chi}_{\rho}(\cdot)\right]^2 = 0.$$

5.2 *Large Deviations and Importance Sampling Theory for Diffusion in Random Environments*

Now that we have defined the random environment and explained its properties, let us review the related large deviations and importance sampling theory from [49, 50]. Set for notational convenience $\tilde{\xi} = D\tilde{\chi}_0$.

Theorem 5.3 (Theorem 3.5 in [49]) *Let $\{(X^{\epsilon,\gamma}, Y^{\epsilon,\gamma}), \epsilon > 0\}$ be, for fixed $\gamma \in \Gamma$, the unique strong solution to (22). Assume non-degeneracy of the diffusion coefficients as well as $\mathcal{C}^1$ smoothness and boundedness of the drift and diffusion coefficients. Consider the regime where $\epsilon, \delta \downarrow 0$ such that $\epsilon/\delta \uparrow \infty$. Then, $\{X^{\epsilon,\gamma}, \epsilon > 0\}$ converges in probability, almost surely with respect to the random environment $\gamma \in \Gamma$, as $\epsilon, \delta \downarrow 0$ to the solution of the ODE*

$$d\bar{X}_t = r(\bar{X}_t)dt$$

and satisfies, almost surely with respect to $\gamma \in \Gamma$, the large deviations principle with rate function

$$S_{t_0T}(\phi) = \begin{cases} \frac{1}{2}\int_{t_0}^{T}(\dot{\phi}_s - r(\phi_s))^T q^{-1}(\phi_s)(\dot{\phi}_s - r(\phi_s))ds & \text{if}\phi \in \mathcal{AC}([t_0, T])\text{and}\phi_{t_0} = x_0 \\ +\infty & \text{otherwise.} \end{cases}$$

where

$$\begin{aligned} r(x) &= \lim_{\rho\downarrow 0}\mathbb{E}^{\pi}\left[\tilde{c}(x,\cdot) + D\tilde{\chi}_{\rho}(\cdot)\tilde{g}(x,\cdot)\right] = \mathbb{E}^{\pi}[\tilde{c}(x,\cdot) + \tilde{\xi}(\cdot)\tilde{g}(x,\cdot)] \\ q(x) &= \lim_{\rho\downarrow 0}\mathbb{E}^{\pi}\left[(\tilde{\sigma}(x,\cdot) + D\tilde{\chi}_{\rho}(\cdot)\tilde{\tau}_1(\cdot))(\tilde{\sigma}(x,\cdot) + D\tilde{\chi}_{\rho}(\cdot)\tilde{\tau}_1(\cdot))^T\right. \\ &\quad \left. + \left(D\tilde{\chi}_{\rho}(\cdot)\tilde{\tau}_2(\cdot)\right)\left(D\tilde{\chi}_{\rho}(\cdot)\tilde{\tau}_2(\cdot)\right)^T\right] \\ &= \mathbb{E}^{\pi}\left[(\tilde{\sigma}(x,\cdot) + \tilde{\xi}(\cdot)\tilde{\tau}_1(\cdot))(\tilde{\sigma}(x,\cdot) + \tilde{\xi}(\cdot)\tilde{\tau}_1(\cdot))^T + \left(\tilde{\xi}(\cdot)\tilde{\tau}_2(\cdot)\right)\left(\tilde{\xi}(\cdot)\tilde{\tau}_2(\cdot)\right)^T\right] \end{aligned}$$

and $\rho = \rho(\epsilon) = \frac{\delta^2}{\epsilon}$.

Notice that the coefficients $r(x)$ and $q(x)$ are obtained by homogenizing (22) by taking $\delta \downarrow 0$ with ϵ fixed. The form of the action functional can be recognized as the one that would come up when considering large deviations for the homogenized system. This is also implied by the fact that δ goes to zero faster than ϵ, since $\epsilon/\delta \uparrow \infty$.

We also remark here that if $b = 0$, then $\chi_\rho = 0$. In this case $r(x), q(x)$ take the simplified forms $r(x) = \mathbb{E}^\pi[\tilde{c}(x,\cdot)]$ and $q(x) = \mathbb{E}^\pi\left[\tilde{\sigma}(x,\cdot)\tilde{\sigma}(x,\cdot)^T\right]$.

Turning now to importance sampling, given controls u_1 and u_2 one considers the controlled dynamics under the importance sampling measure $\bar{\mathbb{P}}$

$$\begin{aligned}
d\bar{X}_s^\epsilon &= \left[\frac{\epsilon}{\delta} b\left(\bar{Y}_s^\epsilon, \gamma\right) + c\left(\bar{X}_s^\epsilon, \bar{Y}_s^\epsilon, \gamma\right) + \sigma\left(\bar{X}_s^\epsilon, \bar{Y}_s^\epsilon, \gamma\right) u_1(s)\right] dt + \sqrt{\epsilon}\sigma\left(\bar{X}_s^\epsilon, \bar{Y}_s^\epsilon, \gamma\right) d\bar{W}_s, \\
d\bar{Y}_s^\epsilon &= \frac{1}{\delta}\left[\frac{\epsilon}{\delta} f\left(\bar{Y}_s^\epsilon, \gamma\right) + g\left(\bar{X}_s^\epsilon, \bar{Y}_s^\epsilon, \gamma\right) + \tau_1\left(\bar{Y}_s^\epsilon, \gamma\right) u_1(s) + \tau_2\left(\bar{Y}_s^\epsilon, \gamma\right) u_2(s)\right] dt \\
&\quad + \frac{\sqrt{\epsilon}}{\delta}\left[\tau_1\left(\bar{Y}_s^\epsilon, \gamma\right) d\bar{W}_s + \tau_2\left(\bar{Y}_s^\epsilon, \gamma\right) d\bar{B}_s\right], \qquad (24) \\
\bar{X}_{t_0}^\epsilon &= x_0, \quad \bar{Y}_{t_0}^\epsilon = y_0
\end{aligned}$$

where $(v_1(s), v_2(s))$ denote the first and second component of the control

$$u(s, \bar{X}_s^\epsilon, \bar{Y}_s^\epsilon) = (u_1(s, \bar{X}_s^\epsilon, \bar{Y}_s^\epsilon), u_2(s, \bar{X}_s^\epsilon, \bar{Y}_s^\epsilon)).$$

Then, for a given cost function $h(x)$, under $\bar{\mathbb{P}}$

$$\Delta^{\epsilon,\gamma}(t_0, x_0, y_0) = \exp\left\{-\frac{1}{\epsilon} h(\bar{X}_T^\epsilon)\right\} \frac{d\mathbb{P}}{d\bar{\mathbb{P}}}(\bar{X}^\epsilon, \bar{Y}^\epsilon),$$

is an unbiased estimator for $\mathbb{E}\left[\exp\left\{-\frac{1}{\epsilon} h(X_T^\epsilon)\right\}\right]$.

Consider next the Hamiltonian

$$H(x,p) = \langle r(x), p\rangle - \frac{1}{2}\left\|q^{1/2}(x)p\right\|^2$$

with $r(x), q(x)$ the coefficients defined in Theorem 5.3 and consider the HJB equation associated to this Hamiltonian, letting $\bar{U}(t,x)$ be a smooth subsolution to it (analogously to Sect. 2 with $r(x)$ and $q(x)$ in place of $-\nabla V(x)$ and Γ, respectively). Then, the following theorem guarantees at least logarithmic asymptotically good performance.

Theorem 5.4 (Theorem 4.1 in [50]) *Let $\{(X_s^\epsilon, Y_s^\epsilon), \epsilon > 0\}$ be the solution to (22) for $s \in [t_0, T]$ with initial point (x_0, y_0) at time t_0. Consider a non-negative, bounded and continuous function $h : \mathbb{R}^m \mapsto \mathbb{R}$. Let $\bar{U}(s,x)$ be a subsolution to the associated HJB equation that has continuous derivatives up to order 1 in t and order 2 in x, and the first and second derivatives in x are uniformly bounded. Assume non-degeneracy of the diffusion coefficients as well $\mathcal{C}^1$ smoothness and boundedness of the drift and*

diffusion coefficients. In the general case where $b \neq 0$*, consider* $\rho > 0$ *and define the (random) feedback control* $u_\rho(s,x,y,\gamma) = \left(u_{1,\rho}(s,x,y,\gamma), u_{2,\rho}(s,x,y,\gamma)\right)$ *by*

$$u_\rho(s,x,y,\gamma) = \left(-\left(\sigma + D\chi_\rho\tau_1\right)^T(x,y,\gamma)\nabla_x\bar{U}(s,x), -\left(D\chi_\rho\tau_2\right)^T(y,\gamma)\nabla_x\bar{U}(s,x)\right)$$

Then for $\rho = \rho(\epsilon) = \frac{\delta^2}{\epsilon} \downarrow 0$ *we have that almost surely in* $\gamma \in \Gamma$

$$\liminf_{\epsilon\to 0} -\epsilon \ln Q^{\epsilon,\gamma}(t_0,x_0,y_0;u_\rho(\cdot)) \geq G(t_0,x_0) + \bar{U}(t_0,x_0). \tag{25}$$

If $b = 0$*, then set* $u(s,x,y,\gamma) = \left(-\sigma^T(x,y,\gamma)\nabla_x\bar{U}(s,x), 0\right)$ *and (25) holds with* $u_\rho(\cdot) = u(\cdot)$.

5.3 A Simulation Study

Consider, for instance, the case of Example 5.1

$$dX_t^{\epsilon,\delta} = -\nabla V^\epsilon\left(X_t^{\epsilon,\delta}, \frac{X_t^{\epsilon,\delta}}{\delta}\right)dt + \sqrt{2\epsilon}dW_t, \tag{26}$$

where the potential function $V^\epsilon(x,x/\delta) = \epsilon Q(x/\delta) + V(x)$. $Q(y)$ is a stationary ergodic random field on a probability space $(\mathcal{X},\mathcal{G},\nu)$. We may consider, for instance, $V(x) = \frac{1}{2}x^2$ and

$$Q(y)\text{meanzeroGaussianwithE}^\nu\left[Q(x)Q(y)\right] = \exp\left[-|x-y|^2\right]$$

Making the connection with (22), the fast Y motion essentially is $Y = X/\delta$. Referring to Theorems 5.3 and 5.4 we have $r(x) = -V'(x)/(K\hat{K})$ and $q = 2/(K\hat{K})$ where $K = \mathrm{E}^\nu[e^{-Q(z)}], \hat{K} = \mathrm{E}^\nu[e^{Q(z)}]$. Given a classical subsolution $\bar{U}$, one expects that the corresponding change of simulation measure that guarantees at least asymptotic optimality is based on the control $\bar{u}(s,x,y,\gamma) = (-\sqrt{2}(1 + \partial\chi(y,\gamma)/\partial y)\bar{U}_x(s,x), 0)$ where one can compute that the weight function is $1 + \partial\chi(y,\gamma)/\partial y = e^{Q(y,\gamma)}/\hat{K}$. Note that in contrast to the periodic case, the control u is random in that it implicitly depends on $\gamma \in \Gamma$, via the random field $Q(y,\gamma)$.

Assume that we want to estimate

$$\theta^{\epsilon,\delta} = \mathrm{P}\left[X^{\epsilon,\delta}\text{hits1before0}|X_0^{\epsilon,\delta} = 0.1\right] \tag{27}$$

As in Sect. 4.1, we compare the asymptotical optimal change of measure with standard Monte Carlo, which corresponds to no change of measure, and with the importance sampling that corresponds to the change of measure based only on the homogenized problem, which ignores the macroscopic problem. Based on 10^7 trajectories, we have the following simulation data

Table 4 Comparing different importance sampling estimators with $x^- = 0$ (equilibrium), $x_0 = 0.1$ (initial point), $x^+ = 1$ (target)

No.	ϵ	δ	ϵ/δ	$\hat{\theta}_1(\epsilon,\delta)$	$\hat{\rho}_0(\epsilon,\delta)$	$\hat{\rho}_1(\epsilon,\delta)$	$\hat{\rho}_2(\epsilon,\delta)$
1	0.25	0.1	2.5	$1.38e-1$	3	0.5	3
2	0.125	0.04	3.125	$1.31e-2$	7	16	8
3	0.0625	0.018	3.472	$6.13e-4$	36	18	42
4	0.05	0.01	5	$2.30e-5$	212	28	316
5	0.04	0.007	5.72	$5.93e-6$	396	75	332
6	0.025	0.004	6.25	$7.82e-10$	–	22	1856

It is clear that the importance sampling scheme based on the asymptotically optimal change of measure $\bar{u}(t,x,y,\gamma)$ outperforms the standard Monte Carlo estimator in which no change of measure is being done. It also outperforms, the estimator based solely on the homogenized system, which ignores the local information characterized by solution to the macroscopic problem. Of course, this behavior is parallel to the behavior observed in the periodic case of Sect. 4.1. Additional simulation studies can be found in [50].

In [50], the interested reader can find further simulation studies in the case of the general model (22) where one does not necessarily have the Y motion to be X/δ. However, we do point out that the theoretical results of [50] are valid for the system (22) where the process (X^ϵ, Y^ϵ) has initial point (x_0, y_0) and both x_0 and y_0 are of order one as $\delta \downarrow 0$. This is not exactly the same to the case where $Y = X/\delta$, as then $y_0 = x_0/\delta$, which is no longer of order one as $\delta \downarrow 0$. But, simulation studies, as the one presented in Table 4, indicate that the theoretical results should be also valid for the $Y = X/\delta$ case.

6 Importance Sampling for Metastable Multiscale Processes and Further Challenges

In Sect. 3 we elaborated on the effects of rest points and metastable dynamics on importance sampling schemes. The end conclusion was that extra care is needed when stable or unstable equilibrium points are in the domain of interest. In this case, asymptotic optimality is not enough in that asymptotically optimal schemes may not perform well in practice unless one goes to really small values of ϵ, in which case the events may be too rare to be of any practical interest. Then, in Sects. 4 and 5 we summarized the issues that come up in the design of asymptotically efficient importance sampling schemes when the dynamics have multiple scales.

In [15, 23] we have systematically addressed the effects of rest points onto the design of importance sampling schemes and have identified what the main issues are. In [23], we have suggested a potential provably appropriate remedy to the issue, by constructions as the ones mentioned in Sect. 3. The subsolution constructed there effectively yields a very good approximation to the zero variance change

of measure. Even though the constructions in [15, 23] work provably well pre-asymptotically and asymptotically and do not degrade as parameters such as the time horizon T getting large, the performance in higher dimensions can be worse than the corresponding performance in the lower-dimensional cases. While this is expected to be the case as the dimension gets larger, due to further approximations and simplifications that need to be made, there is a clear room for improvement here. This is part of ongoing work of the author and we refer the interested reader to [45] for some results in the infinitely dimensional small noise SPDE case.

Moreover, it is clear that the constructions of Sects. 4 and 5 guarantee only asymptotic optimality. If in addition to multiscale dynamics one has to also face metastability, then, as it was seen in Sect. 3, theoretical asymptotic optimality is not sufficient for good numerical performance. One can of course combine the results of Sect. 3 with those of Sects. 4 and 5. To be more precise, one can combine the results of [15, 23] with those of [22, 50]. In practice, one can just use the changes of measure as indicated in [22, 50] that guarantee asymptotic optimality, but construct the subsolution $\bar{U}(t,x)$ as indicated in [15, 23]. We plan to address this issue in more detail in a future work.

Acknowledgements This work was partially supported by the National Science Foundation CAREER award DMS 1550918.

References

1. S. Asmussen, P.W. Glynn, *Stochastic Simulation: Algorithms and Analysis* (Springer, New York, 2007)
2. P. Banushkina, M. Meuwly, Diffusive dynamics on multidimensional rough free energy surfaces. J. Chem. Phys. **127**, 135101 (2007)
3. A. Bensoussan, J.L. Lions, G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, vol. 5, Studies in Mathematics and its Applications (North-Holland Publishing Co., Amsterdam, 1978)
4. J.H. Blanchet, P. Glynn, Efficient rare-event simulation for the maximum of heavy-tailed random walks. Ann. Appl. Prob. **18**, 1351–1378 (2008)
5. J.H. Blanchet, J.C. Liu, State-dependent importance sampling for regularly varying random walks. Adv. Appl. Probab. **40**, 1104–1128 (2008)
6. J.H. Blanchet, P. Glynn, J.C. Liu, Fluid heuristics, Lyapunov bounds and efficient importance sampling for a heavy-tailed G/G/1 queue. Queueing Syst. **57**, 99–113 (2007)
7. J.H. Blanchet, P. Glynn, K. Leder, On Lyapunov inequalities and subsolutions for efficient importance sampling. ACM TOMACS **22**(3), Artical No. 13 (2012)
8. A. Bovier, M. Eckhoff, V. Gayrard, M. Klein, Metastability in reversible diffusion processes I. Sharp estimates for capacities and exit times. J. Eur. Math. Soc. **6**, 399–424 (2004)
9. A. Bovier, V. Gayrard, M. Klein, Metastability in reversible diffusion processes II. Precise estimates for small eigenvalues. J. Eur. Math. Soc. **7**, 69–99 (2005)
10. P. Boyle, M. Broadie, P. Glasserman, Monte Carlo methods for security pricing. J. Econ. Dyn. Control. **21**, 1257–1321 (1997)
11. M. Cameron, Finding the quasipotential for nongradient SDEs. Phys. D: Nonlinear Phenom. **241**(18), 1532–1550 (2012)

12. M. Day, T. Darden, Some regularity results on the Ventcel-Freidlin quasi-potential function. Appl. Math Opt. **13** 259–282 (1985)
13. A. Dembo, O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 38, 2nd ed., Applications of Mathematics (Springer, New York, 1998)
14. P. Dupuis, K. Spiliopoulos, Large deviations for multiscale problems via weak convergence methods. Stochastic Process. Appl. **122**, 1947–1987 (2012)
15. P. Dupuis, K. Spiliopoulos, Rare event simulation in the neighborhood of a rest point, in *2014 Winter Simulation Conference* (IEEE, 2014), pp. 564–573
16. P. Dupuis, H. Wang, Importance sampling, large deviations and differential games. Stochastics Stochastics Rep. **76**, 481–508 (2004)
17. P. Dupuis, H. Wang, Subsolutions of an Isaacs equation and efficient schemes of importance sampling. Math. Oper. Res. **32**, 723–757 (2007)
18. P. Dupuis, K. Leder, H. Wang, Large deviations and importance sampling for a tandem network with slow-down. Queueing Syst. **57**, 71–83 (2007)
19. P. Dupuis, A. Sezer, H. Wang, Dynamic importance sampling for queueing networks. Ann. Appl. Probab. **17**, 1306–1346 (2007)
20. P. Dupuis, K. Leder, H. Wang, Importance sampling for weighted serve-the-longest-queue. Math. Oper. Res. **34**(3), 642–660 (2009)
21. P. Dupuis, K. Spiliopoulos, H. Wang, Rare event simulation in rough energy landscapes, in *2011 Winter Simulation Conference* (2011), pp. 504–515
22. P. Dupuis, K. Spiliopoulos, H. Wang, Importance sampling for multiscale diffusions. Multiscale Model. Simul. **12**(1), 1–27 (2012)
23. P. Dupuis, K. Spiliopoulos, X. Zhou, Escape from an equilibrium: importance sampling and rest points I. Ann. Appl. Probab. **25**(5), 2909–2958 (2015)
24. W.H. Fleming, M.R. James, Asymptotic series and exit time probabilities. Ann. Probab. **20**(3), 1369–1384 (1992)
25. W.H. Fleming, H.M. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd edn. (Springer, Berlin, 2006)
26. M. Freidlin, R. Sowers, A comparison of homogenization and large deviations with applications to wavefront propagation. Stochastic Process Appl. **82**, 23–52 (1999)
27. M.I. Freidlin, A.D. Wentzell, *Random Perturbations of Dynamical Systems*, 2nd edn. (Springer, New York, 1988)
28. P. Glasserman, *Monte Carlo Methods in Financial Engineering* (Springer, New York, 2004)
29. P. Glasserman, S. Kou, Analysis of an important sampling estimator for tandem queues. ACM Trans. Model. Comput. Simul. **4**, 22–42 (1995)
30. P. Glasserman, Y. Wang, Counter examples in importance sampling for large deviations probabilities. Ann. Appl. Probab. **7**, 731–746 (1997)
31. P.W. Glynn, D.L. Iglehart, Simulation methods for queues: an overview. Queueing Syst.: Theory Appl. **3**, 221–256 (1988)
32. R.C. Griffiths, S. Tavaré, Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46**, 131–159 (1994)
33. W. Janke, *Rugged Free-Energy Landscapes*, Lecture Notes in Physics, vol. 736/2008 (Springer, Berlin, 2008)
34. T. Komorowski, C. Landim, S. Olla, *Fluctuations in Markov Processes: Time Symmetry and Martingale Approximation* (Springer, Berlin, 2012)
35. E. Kosygina, F. Rezakhanlou, S.R.S. Varadhan, Stochastic homogenization of Hamilton-Jacobi-Bellman equations, Commun. Pure Appl. Math. **LIX**, 0001–0033 (2006)
36. R.D. Levine, Monte Carlo, maximum entropy and importance sampling. Chem. Phys. **228**, 255–264
37. S. Lifson, J.L. Jackson, On the self-diffusion of ions in a polyelectrolyte solution. J. Chem. Phys. **36**, 2410–2414 (1962)
38. R.S. Maier, D.L. Stein, Escape problem for irreversible systems. Phys. Rev. E **48**(2), 931–938 (1993)

39. R.S. Maier, D.L. Stein, Limiting exit location distributions in the stochastic exit problem. SIAM J. Appl. Math. **57**(3), 752–790 (1997)
40. O. Mazonka, C. Jarzynski, J. Blocki, Computing probabilities of very rare events for Langevin processes: a new method based on importance sampling. Nucl. Phys. A **641**, 335–354 (1998)
41. S. Olla, Homogenization of diffusion processes in random fields (1994). Available at www.ceremade.dauphine.fr/~olla/lho.ps
42. H. Osada, Homogenization of diffusion processes with random stationary coefficients, in *Probability Theory and Mathematical Statistics*. Lecture Notes in Mathematics, vol. 1021 (Springer, Berlin, 1983), pp. 507–517
43. G.A. Pavliotis, A.M. Stuart, Multiscale Methods: Averaging and Homogenization (Springer, Berlin, 2007)
44. G. Papanicolaou, S.R.S. Varadhan, Boundary value problems with rapidly oscillating random coefficients, in *Colloquia Mathematica Societatis Janos Bolyai 27, Random Fields*, Esztergom (Hungary) 1979, North Holland (1982), pp. 835–873
45. M. Salins, K. Spiliopoulos, Rare event simulation via importance sampling for linear SPDE's. stochostics and Partial Differential Equation: Analysis and computations (accepted, 2017)
46. D. Siegmund, Importance sampling in the Monte Carlo study of sequential tests. Ann. Stat. **4**, 673–684 (1976)
47. K. Spiliopoulos, Large deviations and importance sampling for systems of slow-fast motion. Appl. Math. Optim. **67**, 123–161 (2013)
48. K. Spiliopoulos, Non-asymptotic performance analysis of importance sampling schemes for small noise diffusions. J. Appl. Probab. **52**, 1–14 (2015)
49. K. Spiliopoulos, Quenched large deviations for multiscale diffusion processes in random environments. Electron. J. Probab. **20**(15), 1–29 (2015)
50. K. Spiliopoulos, Rare event simulation for multiscale diffusions in random environments. SIAM Multiscale Model. Simul. **13**(4), 1290–1311 (2015)
51. E. Vanden-Eijnden, J. Weare, Rare event simulation with vanishing error for small noise diffusions. Commun. Pure Appl. Math. **65**(12), 1770–1803 (2012)
52. S.R.S. Varadhan, *Large Deviations and Applications*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 46 (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1984)
53. A. Viel, M.V. Patel, P. Niyaz, K. Whaley, Importance sampling in rigid body diffusion Monte Carlo. Comput. Phys. Commun. **145**, 24–47 (2002)
54. D. Zuckerman, T. Woolf, Efficient dynamic importance sampling of rare events in one dimension. Phys. Rev. E **63**(016702), 1–10 (2000)
55. R. Zwanzig, Diffusion in a rough potential. Proc. Natl. Acad. Sci. U. S. A. **85**, 2029–2030 (1988)

Multiscale Simulation of Stochastic Reaction-Diffusion Networks

Stefan Engblom, Andreas Hellander, and Per Lötstedt

2010 *Mathematics Subject Classification*. Primary: 65C40; Secondary: 60H35; Tertiary: 92C05

1 Introduction

A defining theme in molecular systems biology has been the realization that stochastic effects on biochemical networks are important factors to consider when studying the function of molecular control systems. As a consequence, discrete stochastic models are more adequate than deterministic models because of the inherent discreteness in chemical systems when the copy numbers of the molecular species are small [11, 31, 66, 91, 105, 108, 110, 115, 119, 121, 138, 146].

Biochemical network model simulation offers many challenges due to their inherent complexity, randomness, and the presence of time scale separation. While to date, most models are relatively small in the number of molecular components due to the formidable challenge to parametrize models, a milestone whole cell simulation was recently conducted including all biochemical subsystems of a simple bacterium, *Mycoplasma genitalium* [88]. Such a simulation has been characterized as the grand challenge of the twenty-first century in [141]. Despite the vast parameter space and many subsystems, a cell cycle is simulated in [88] and conclusions can be drawn from genotype to phenotype. Good agreement with experiments is obtained and predictions are made that can be validated against experiments.

S. Engblom • A. Hellander • P. Lötstedt (✉)
Division of Scientific Computing, Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden
e-mail: stefane@it.uu.se; andreash@it.uu.se; perl@it.uu.se

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_3

A great challenge for simulations of increasingly complex models is the large separation in kinetics rate constants, diffusion constants, and molecular copy numbers typically observed in network models. If realistic cellular geometries are considered with molecules moving in 3D and binding to 2D and 1D structures, the situation becomes even more complicated. All these reasons call for multiscale methods where the most efficient simulation methodology that is able to resolve the local model features with sufficient accuracy is used to arrive at a feasible simulation. We here review the most commonly employed models on a *microscopic*, *mesoscopic*, and *macroscopic* scale, as well as the state of the art of methods and algorithms for computer simulation of deterministic and stochastic biochemical systems with chemical reactions and a spatial variation of the molecules.

Stochastic models for chemical reactions are found in [54, 106, 142]. In a well-stirred system with strong diffusive mixing, the space dependence can be ignored. This is not always the case and examples in [42, 97] show that diffusion is important for the dynamics of the system. Other recent review articles on this subject are [14, 22, 40, 62, 63, 103, 118, 130].

The different levels of description can be viewed as models on different levels of fidelity, ascending from a detailed, stochastic, and low level, to an intermediate stochastic level, and to a deterministic high level of modeling. A short description of the modeling levels is as follows.

Microscopic Level A molecule is assumed to be spherical and moves by Brownian motion in an *off-lattice* model. Molecules react with a certain probability when they are in the neighborhood of each other. Since a molecule occupies a part of space, effects of volume exclusion or crowding can be simulated. The trajectories of all molecules in the system are followed separately in a Monte Carlo simulation in an accurate but computationally expensive model. Many realizations of the system are necessary to obtain accurate statistics of the behavior. An overview of this model is found in Sect. 2.

Mesoscopic Level The volume Ω is partitioned into voxels $\mathcal{V}_j$, $j = 1, \ldots, N$, covering Ω without overlap in this *on-lattice* model. The copy numbers of the species in each voxel define the state of the system. The state is changed by reactions between the molecules in a voxel and by jump events where a molecule in one voxel moves to a neighboring voxel in a continuous time Markov process. There is a master equation for the probability distribution of the time dependent state of the system which can be solved analytically and numerically for a few simple models. Otherwise, sample paths of the system are generated via Monte Carlo simulations. This model is described in more detail in Sect. 3.

Macroscopic Level The concentrations of the chemical species satisfy partial differential equations (PDEs) in the macroscopic model. The PDEs are equations for the time evolution of the concentrations with a diffusion term and source terms given by the chemical reactions. If the diffusion is so efficient that the spatial variation is negligible, then the equations for the concentrations are simplified to a system of ordinary differential equations (ODEs). Compared to the other two models, this model is computationally inexpensive but less accurate. The model is applicable in systems with many molecules and a large Ω and is treated in Sect. 4.

The notation is as follows. The biochemical system in a volume Ω has M molecular species $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$ reacting with each other in R chemical reactions defined by ν_{rm} and ν'_{rm} for $r = 1, \ldots, R$, $m = 1, \ldots, M$. In reaction r, a sum of ν_{rm} molecules of species m react and form a sum of ν'_{rm} molecules of species n. A reaction can be written

$$\sum_{m=1}^{M} \nu_{rm} S_m \longrightarrow \sum_{n=1}^{M} \nu'_{rm} S_n, \ r = 1, \ldots, R. \tag{1}$$

The set of all reaction channels is $\mathcal{R} = \{\boldsymbol{\nu}_r \to \boldsymbol{\nu}'_r, \ r = 1, \ldots, R\}$. The propensity for the reaction to occur per time unit is a_r and depends on the concentrations of the species and may depend on the position $\mathbf{r} \in \Omega$. The molecules are here assumed to be transported in Ω by diffusion, but molecules can also move by active transport in certain directions, for example, along actin filaments and microtubules. A vector $\mathbf{u}$ has the components u_i and the elements of a matrix $\mathbf{A}$ are A_{ij}. The norm $\|\cdot\|$ is the Euclidean vector norm and its subordinate spectral matrix norm.

2 The Microscopic Scale

On the microscopic level, individual molecules are tracked as they diffuse and react in a given domain. Two formalisms have received particular attention in the molecular systems biology area: the Doi-model [23, 82] and the Smoluchowski diffusion-limited reaction (SDLR) model [144]. In both cases, the change in the positions due to diffusion is governed by simple Brownian motion.

In the Doi-type model, two molecules react with a uniform rate whenever their separation is less than a stipulated reaction radius, while in the SDLR model, reactions occur with a fixed rate whenever the molecule are separated by precisely the reaction radius. Hence, in the former reactions are treated as a reaction potential, whereas in the latter they are treated as a boundary condition at the contact point between molecules. The relationship between the models is discussed in greater detail in [1] with rigorous convergence results for some important special cases. In what follows we will mostly be concerned with the SDLR model.

2.1 Smoluchowski Diffusion-Limited Kinetics

Assume that a spherical molecule of species X is at position $\mathbf{r_0}$ at time t_0. The probability $p(\mathbf{r}, t) = p(\mathbf{r}, t|\mathbf{r_0}, t_0)$ to find a molecule at position $\mathbf{r}$ in free space after time $t > t_0$ is the solution to the diffusion equation

$$\begin{aligned} \frac{\partial p(\mathbf{r}, t)}{\partial t} &= \gamma_X \Delta_r p(\mathbf{r}, t), \\ p(\mathbf{r}, t_0) &= \delta(\mathbf{r} - \mathbf{r}_0), \ \lim_{\|\mathbf{r}\| \to \infty} p(\mathbf{r}, t) = 0. \end{aligned} \tag{2}$$

The initial condition at $t = t_0$ is the Dirac delta and γ_X is a diffusion constant. The solution to (2) is the Green's function for the diffusion equation.

When the X-molecule is involved in a bimolecular reaction with species Y in free space then the probability distribution for the distance $\mathbf{r}$ between the two molecules satisfies

$$\begin{aligned} \frac{\partial p(\mathbf{r},t)}{\partial t} &= D\Delta_r p(\mathbf{r},t), \\ 4\pi\sigma^2 D\frac{\partial p}{\partial r}|_{\|\mathbf{r}\|=\sigma} &= k_r p(\|\mathbf{r}\| = \sigma, t), \\ p(\mathbf{r},t_0) &= \delta(\mathbf{r}-\mathbf{r}_0), \lim_{\|\mathbf{r}\|\to\infty} p(\mathbf{r},t) = 0. \end{aligned} \tag{3}$$

Here σ is the sum of the reaction radii of the hard-sphere species X and Y with radius σ_X and σ_Y, respectively, and D is the sum of the diffusion constants γ_X and γ_Y. The derivative in the boundary condition is taken in the radial direction at $\|\mathbf{r}\| = \sigma$. The probability density for the common center of the two molecules obeys a diffusion equation as in (2), see [143]. The bimolecular chemical reaction is modeled by the Robin boundary condition, where the rate of change of the probability of reaction at the contact point between molecules is proportional to the intrinsic, or microscopic, reaction rate constant k_r. This parameter is in practical modeling normally considered to be a given parameter, but it can also be related to other fundamental constants and the probability of reaction given a molecular encounter [61, Chap. 4.8].

2.2 *Software for Microscale Simulation*

Software packages for particle-based simulation can roughly be divided into two categories: approximate Brownian Dynamics where the microscopic model is approximated in some formal sense, relying on discretized time or space, and those relying on a mathematically systematic approximation to (2) and (3). The former category includes Smoldyn [4, 5] and MCell [90]. Smoldyn relies on continuous space with discretized time, while the latter uses a number of discrete spatial directions in which the molecule can move. Spatiocyte [147] uses a microscopic molecule-sized hexagonal lattice and updates the system in a rule-based manner. For a recent and more comprehensive review of particle-based simulation tools, see [130]. Brownian Dynamics becomes expensive due to the need to take small time steps to resolve molecule collisions with discretized time but naturally accommodate volume exclusion effects at high molecular densities. At low to intermediary densities, Green's function reaction dynamics can be a more efficient alternative.

2.3 Green's Function Reaction Dynamics

Simulation of the SDLR model by solving (2) and (3) and using the solution to sample particle positions and reaction events is relatively straightforward for a single molecule pair, but quickly becomes intractable for multiple molecules and reactions. In Green's function reaction dynamics (GFRD) [143] this problem is addressed by grouping molecules and restricting the time step such that it is unlikely that a molecule react with a molecule other than its closest neighbor. In this way, the many-body problem is simplified to solving many two-body problems over a shorter time window. The method is implemented using analytical solutions to (2) and (3), tabulated values or for increased flexibility with operator splitting and numerical solutions [71]. For relatively dilute systems, the GFRD methodology can result in much faster simulations for the same accuracy level compared to more traditional reactive Brownian dynamics codes [143], especially if the first-passage time Kinetic Monte Carlo (fpKMC) [24, 117] algorithm is employed, such as in the eGFRD software [136].

In many situations, especially if reactions are not too diffusion limited or if there are higher copy number counts of most species, a mesoscopic model in which space is discretized and the number of particles at each location in space is only tracked up to the chosen spatial resolution can offer far more efficient simulations than the microscopic methods.

3 The Mesoscopic Scale

Below we summarize the physical and mathematical background for mesoscopic stochastic modeling of chemical reactions. The very term "mesoscopic" suggests the level somewhere *between* the microscopic and the macroscopic levels of description. The state variable is the same as in the latter (number of molecules or concentration), but randomness has been introduced in order to better capture the microscopic conditions.

A traditional derivation of the chemical master equation (CME) from micro-physical assumptions is given in Sect. 3.1. Direct solution of the CME is chiefly discussed in Sect. 3.2 and in Sect. 3.3. The spatially inhomogeneous case is introduced in the form of a general modeling formalism. Different types of sampling algorithms are summarized in Sect. 3.4 and consistency relations with microscopic models in the diffusion-limited context are discussed in Sect. 3.5.

3.1 Derivation of the CME

Derivations from first principles of mesoscopic models of well-stirred chemical reactions are found in [32, 57, 58]. The typical assumption is that of an ideal gas and this is essentially the path we take here, adapted from [34]. See also the monographs [53, 142] which treat general Markovian systems in physics.

To be concrete we shall consider a reactive system of X- and Y-molecules moving around in a vessel Ω of total volume $|\Omega|$. We assume that a single bimolecular reaction occurs

$$X + Y \to \emptyset. \tag{4}$$

In a probabilistic description of the system, let $x(t)$ and $y(t)$ count the two kinds of molecules at time t. Let C be the event that a randomly sampled pair of molecules collides in the interval of time $[t, t + dt)$ and let $V(\mathbf{v})$ be the event that this pair of molecules has relative velocity $\mathbf{v} = \mathbf{v}_X - \mathbf{v}_Y$. $P(C)$ may now be expanded as a conditional probability,

$$P(C) = \int_{\mathbf{v}} P(C|V(\mathbf{v}))P(V(\mathbf{v})), \tag{5}$$

with $P(C|V(\mathbf{v}))$ the probability that two randomly selected molecules collide given that they have relative velocity $\mathbf{v}$.

1. Assuming that the system of molecules is homogeneous, the probability of finding any randomly selected molecule inside some volume $\Delta\Omega$ is just $|\Delta\Omega|/|\Omega|$. Hence we deduce

$$P(C|V(\mathbf{v})) = \frac{\rho(v, dt)}{|\Omega|}, \tag{6}$$

   where $\rho(v, dt)$ is the volume of the region in which an X-molecule with speed $v = \|\mathbf{v}\|$ relative to a Y-molecule lies, given that the pair is to collide in $[t, t+dt)$. This region is an extruded shape of length $v\,dt$ and obeys the scaling law

$$\rho(\alpha v, \beta dt) = \alpha\beta\rho(v, dt). \tag{7}$$

   It follows from (7) that the conditional collision probability in fact simplifies into

$$P(C|V(\mathbf{v})) = \frac{\rho v\,dt}{|\Omega|}, \tag{8}$$

   where the constant ρ depends on the precise shapes of the molecules.
2. In addition to the homogeneity of the solvent, we assume also that the system is in thermal equilibrium. This means that the probability distribution over the velocity is time-independent,

$$P(V(\mathbf{v})) = P_{MB}(\mathbf{v})d^3\mathbf{v}, \tag{9}$$

   where the Maxwell–Boltzmann distribution is the usually assumed stationary distribution.

Combining (5), (8), and (9) we find that

$$P(C) = \int_{\mathbf{v}} P_{MB}(\mathbf{v}) \frac{\rho v \, dt}{|\Omega|} \, d^3\mathbf{v} = \frac{\rho E[\|\mathbf{v}\|]}{|\Omega|} \, dt. \tag{10}$$

The main importance of this finding is the linear scaling with dt as this implies the existence of a describing Markov chain.

We now wish to determine the probability that *exactly one* pair of molecules collides in $[t, t + dt)$. The total possible number of colliding pairs is just $x(t)y(t)$. By homogeneity, the event that only one of these pairs collides is formed from a total of xy independent events and the probability for this is

$$k/|\Omega| \, dt(1 - k/|\Omega| \, dt)^{xy-1} = k/|\Omega| \, dt + o(dt), \tag{11}$$

where $k \equiv \rho E[\|\mathbf{v}\|]$ from (10). These events are mutually exclusive and the probability for a single collision is therefore obtained directly by summation. Moreover, the probability that $n > 1$ reactions take place is $O(dt^n) = o(dt)$. Hence we have the probability that in $[t, t + dt)$,

- *exactly one* reaction takes place is $k\,x(t)y(t)/|\Omega| \, dt + o(dt)$;
- *more than one* reaction takes place is $o(dt)$;
- *no* reaction occurs is $1 - k\,x(t)y(t)/|\Omega| \, dt + o(dt)$.

To fully characterize the system, we use the state vector $z(t) = [x(t) \;\; y(t)]^T$, counting the two kinds of molecules at time t. Define also the bimolecular reaction propensity

$$a(z_t) = k_a z_{1t} z_{2t} = kx(t)y(t)/|\Omega|. \tag{12}$$

Let $z(0)$ be a given number of molecules at time $t = 0$ and let $P(z, t|z(0))$ be the conditional probability for a certain state z at time t given this initial state. We claim that

$$\begin{aligned} P(z, t + dt|z(0)) = {} & P(z + [1; 1], t|z(0)) \times [a(z + [1; 1]) \, dt + o(dt)] \\ & + o(dt) + P(z, t|z(0)) \times [1 - a(z) \, dt + o(dt)] \, . \end{aligned} \tag{13}$$

The first term is the probability of the state being at $z + [1; 1]$ multiplied by the probability that the reaction occurs. The second term is the vanishing probability of more than one reaction occurring and the third term is the probability of already being in state z and remaining there.

Taking limits in (13) and suppressing the dependence on the initial state we find

$$\frac{\partial}{\partial t} P(z, t) = a(z + [1; 1])P(z + [1; 1], t) - a(z)P(z, t). \tag{14}$$

This, finally, is the *master equation* for the well-stirred interpretation of the chemical system (4).

For the general system governed by (1), we have a state $x \in \mathbf{Z}_+^M$ and we use the *notation*

$$x \xrightarrow{a_r(x)} x + \nu_r' - \nu_r, \quad r = 1, \ldots, R, \tag{15}$$

to mean that the probability that the state x at time t turns into the new state $x+\nu_r'-\nu_r$ at time $t + dt$ is $a_r(x)\,dt + o(dt)$.

Recall that the *Markov*, or *memoryless property*, means that the conditional probability for the state (x_n, t_n) given the system's full history satisfies

$$P\left(x_n, t_n | x_{n-1}, t_{n-1};\, \ldots;\, x_1, t_1\right) = P\left(x_n, t_n | x_{n-1}, t_{n-1}\right), \tag{16}$$

that is, previous states $(x_{n-2}, t_{n-2}, \ldots, x_1, t_1)$ are remembered only via the last state (x_{n-1}, t_{n-1}). A direct consequence of the Markov property is the *Chapman–Kolmogorov equation*,

$$P(x_3, t_3 | x_1, t_1) = \sum_y P(x_3, t_3 | y, t_2) P(y, t_2 | x_1, t_1). \tag{17}$$

Given an initial state $x(0)$, the time derivative of the conditional probability is given by, using (17),

$$\frac{\partial}{\partial t} P(x, t | x(0)) = \lim_{\Delta t \to 0} \frac{\sum_y P(x, t + \Delta t | y, t) P(y, t | x(0)) - P(y, t + \Delta t | x, t) P(x, t | x(0))}{\Delta t}. \tag{18}$$

Taking limits and using the transition model (15) we obtain in analogy with the reasoning that leads to (14)

$$\frac{\partial}{\partial t} P(x, t) = \sum_{r=1}^{R} a_r(x - \nu_r' + \nu_r) P(x - \nu_r' + \nu_r, t) - a_r(x) P(x, t), \tag{19}$$

again after suppressing the dependency on the initial state. In conclusion, the master equation (19) is a differential form of the Chapman–Kolmogorov equation (17), in turn a direct consequence of the fundamental Markov property (16) under the specific choice of transition model (15).

3.2 Solution of the CME

There are only a few analytical solutions to the CME [51, 85], typically when the propensity a_r is linear in x. The difficulty with straightforward numerical solution of the CME in (19) is the exponential growth in the number of differential equations for P when the number of species M grows ("the curse of dimensionality"). Based on

the premise that "smooth" solutions to the CME can be effectively approximated, a number of attempts have been made at solving the CME using an appropriate basis. Examples include spectral methods [21, 35], a wavelet method [84], sparse grids [69], projection-type methods [114], and sums of low dimensional tensors [89]. For sufficiently high accuracy demands and not too high dimensionality, this is more efficient than repeated Monte Carlo simulations as in Sect. 3.4.

All the moments of the distribution $P(x,t)$ in the CME satisfy equations involving one moment higher thus resulting in a moment closure problem. Assuming that the third moments are negligible, a closed system of ODEs is obtained for the first and second moments [33, 45, 50, 64, 129, 142]. The size of the system is then $M + M(M+1)/2$ for the mean and the covariances avoiding the exponential dependence of M in the CME. A special case is the linear noise approximation [29, 45, 65, 142] for the first two moments where the mean values satisfy the reaction rate equations, see Sect. 4.

3.3 The Reaction-Diffusion Master Equation

The CME fundamentally assumes that the solvent of molecules is homogeneous in space and in thermodynamic equilibrium. There are several interesting situations when those assumptions are broken. For example, when the molecular transport is slow compared to the reaction intensity such that concentration gradients may build up. Also, inside biological cells many reactive processes are localized and the assumption of being well mixed no longer holds. A reasonable idea is to subdivide the domain $|\Omega|$ into smaller computational voxels $\mathcal{V}_j$ such that their individual volume size $|\mathcal{V}_j|$ is sufficiently small to make them behave as approximately well-stirred by the presence of diffusion.

As before we assume that there are M chemically active species X_{ij} for $i = 1,\ldots,M$, but now counted separately in the N introduced voxels, $j = 1,\ldots,N$. It follows that the state of the system can be represented by an array $\mathbf{x}$ with $M \times N$ elements. The jth column of $\mathbf{x}$ is denoted by $\mathbf{x}_{\cdot j}$ and the ith row by $\mathbf{x}_{i\cdot}$. The state $\mathbf{x}$ is now changed by chemical reactions occurring between the molecules in the same voxel *and* by diffusion where molecules move to adjacent voxels. Reactions take place vertically in $\mathbf{x}$ and diffusion is horizontal.

Assuming each voxel to be approximately well-stirred the master equation (19) is valid as a description of the reactions,

$$\begin{aligned}\frac{\partial p(\mathbf{x},t)}{\partial t} &= \mathcal{M}p(\mathbf{x},t)\\ &:= \sum_{j=1}^{N}\sum_{r=1}^{R} a_r(\mathbf{x}_{\cdot j} - \boldsymbol{\nu}_r' + \boldsymbol{\nu}_r)p(\mathbf{x}_{\cdot 1},\ldots,\mathbf{x}_{\cdot j} - \boldsymbol{\nu}_r' + \boldsymbol{\nu}_r,\ldots,\mathbf{x}_{\cdot N},t)\\ &\quad - \sum_{j=1}^{N}\sum_{r=1}^{R} a_r(\mathbf{x}_{\cdot j})p(\mathbf{x},t), \end{aligned} \tag{20}$$

where, for brevity, $p(\mathbf{x},t) = P(\mathbf{x},t|\mathbf{x}(0))$.

A natural type of transition for modeling diffusion from one voxel $\mathcal{V}_k$ to another voxel $\mathcal{V}_j$ is the linear rate law

$$\mathbf{x}_{ik} \xrightarrow{q_{kj}\mathbf{x}_{ik}} \mathbf{x}_{ij}. \tag{21}$$

It is understood in (21) that q_{kj} is non-zero only for those voxels that are connected. The rate q_{kj} should ideally be taken as the inverse of the mean first exit time for a single molecule of species i from voxel $\mathcal{V}_k$ to $\mathcal{V}_j$. By the properties of Brownian motion, $q_{kj} \propto \gamma/h_{kj}^2$, where γ is the diffusion constant, h_{kj} is a measure of the length scale of the voxels, and where the constant of proportionality depends on the precise shape of $\mathcal{V}_k$ [102].

The diffusion master equation can now be written in the same way as (20),

$$\begin{aligned}\frac{\partial p(\mathbf{x},t)}{\partial t} &= \sum_{i=1}^{M}\sum_{k=1}^{N}\sum_{j=1}^{N} q_{kj}(\mathbf{x}_{ik}+\mathbb{D}_{kj,k})p(\mathbf{x}_{1\cdot},\ldots,\mathbf{x}_{i\cdot}+\mathbb{D}_{kj},\ldots,\mathbf{x}_{M\cdot},t)\\ &\quad -q_{kj}\mathbf{x}_{ik}p(\mathbf{x},t) =: \mathcal{D}p(\mathbf{x},t).\end{aligned} \tag{22}$$

The transition vector $\mathbb{D}_{kj,\cdot}$ is zero except for two components: $\mathbb{D}_{kj,k} = 1$ and $\mathbb{D}_{kj,j} = -1$.

By combining (20) and (22), we finally arrive at the *reaction-diffusion master equation* (RDME),

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = (\mathcal{M}+\mathcal{D})p(\mathbf{x},t). \tag{23}$$

The numerical methods for solving the CME in the beginning of Sect. 3.2 are hardly applicable to the RDME since the dimension of $\mathbf{x}$ in (23) is MN. An exception is the finite state projection method in [25]. In general, we have to resort to the Monte Carlo methods reviewed in the next section.

3.4 Simulation Algorithms

Due to the exponential growth in computing time and memory for direct numerical solution of the CME in (19), Monte Carlo methods of less computational complexity are a more efficient alternative to determine the statistical moments of the copy numbers of the species and for simulation of the chemical networks. The first presentation of simulation algorithms with applications in chemical kinetics is due to Gillespie [56], although the algorithm of simulating continuous time Markov chains actually dates back to much earlier work by Doob in the 1950s.

The most straightforward algorithm, the *direct method*, or usually just "the" Stochastic Simulation Algorithm (SSA), continuously determines *what* is the next event and *when* does it happen. The time for the next reaction is first sampled from

an exponential distribution. Which reaction channel to fire is sampled based on the reaction propensities. The state of the chemical system is then updated according to (15) or (21). An SSA with improved efficiency is found, e.g., in [55]. SSA can be recognized as a kind of a signature for explicit time-stepping methods and therefore suffers drawbacks when stiff problems are considered with a large span of inherent time scales; here a slow dynamics (of interest) is masked by rapid fluctuations and transients (of less interest). Methods specifically designed for this situation with two separate time scales are found in [15, 28].

Monte Carlo simulation with SSA using many trajectories for accurate estimates of the moments is also time consuming. Let $\mathbf{X}$ be the (stochastic) state variable for the system. Then it is shown in [96] that $\mathbf{X}(t)$ can be written

$$\mathbf{X}(t) = \mathbf{X}(0) + \sum_{r=1}^{R} Y_r \left(\int_0^t a_r(\mathbf{X}(s))\, ds \right) (\boldsymbol{\nu}_r' - \boldsymbol{\nu}_r), \ t > 0. \tag{24}$$

The state is updated with the stoichiometric vector $\boldsymbol{\nu}_r' - \boldsymbol{\nu}_r$ multiplied by independent, unit rate Poisson processes Y_r depending on the integral of the propensity a_r of the reaction r. By choosing t as a small time step Δt, approximations of the integral in (24) are possible. These approximations are then applied iteratively in a time-stepping process. Time discretization methods, most importantly variants of the Euler forward method, or the so-called *tau leaping methods*, were devised early on for (24) [16, 59, 122, 140]. Such methods gain efficiency by offering a less detailed simulation over longer times, but also suffers from stiffness issues and the risk of having negative copy numbers. Recent further algorithmic improvements and analysis with a switch between tau leaping and SSA are found in [112] and convergence acceleration for expected values is obtained by the Multilevel Monte Carlo method in [2].

Early approaches to handling spatial problems are found in [83, 133, 135], and an effective spatial simulation method, the *Next Subvolume Method* (NSM) was invented in [30] for Cartesian meshes. For general geometries, the work in [38] connected traditional numerical discretization methods with the required diffusion rates such that a consistent modeling framework was possible. Adaptive Cartesian meshes are considered in [10] for simulation of systems with strong spatial gradients.

Freely available spatial simulation software has since been released [26, 68, 77, 101]. Parallel algorithms for spatial problems are found in [6, 9], and is also ongoing research.

3.5 Relationship Between the Microscopic and Mesoscopic Models

The mesoscopic RDME (23) is not formally an approximation of any microscopic model, but rather a model in its own right. It is more efficient than corresponding microscopic simulations when the system permits a coarser approximation and

hence a reasonably coarse mesh (relative to the molecule sizes). However, it is of practical interest to ask to what extent it approximates certain microscopic models as the mesh is refined. The key to defining the relationship between the RDME and the SDLR model is the mesoscopic bimolecular reaction rate k_a in (12), which in most mesoscopic models to date is treated as a given modeling parameter. Given the microscale parameters defined in Sect. 2.1, seminal work by Collins and Kimball defines the classic relationship between the microscopic reaction rate k_r in (3) and the mesoscopic rate $k_a = k/|\Omega|$ in (12), where k is given by

$$k = \frac{4\pi\sigma D k_r}{4\pi\sigma D + k_r}. \tag{25}$$

This relation is valid in 3D and only if the reaction volume $|\Omega|$ is much larger than the molecules, which for the RDME translates to the guidelines $h \gg \sigma$ for the mesh size h. This relationship was also derived more recently from a different perspective in [60].

However, simply using the diffusional propensity function (25) does not guarantee convergence to, or even a good approximation of, the SDLR model as $h \to 0$. Indeed, it was shown by Isaacson that if we let h tend to zero with a fixed reaction radius σ of the molecules, the rate of bimolecular reactions will successively diminish to completely disappear in the limit [81]. Physically this makes sense as a result of the point particle assumption in the RDME becoming invalid on small spatial scales. In fact, there are fundamental limits to the smallest resolution h^* in RDME at which any choice of mesoscale k_a lets even the mean binding time between two molecules match [74]. Expressed in terms of the reaction radii of the involved molecules, these critical limits are given by

$$h^* \approx 5.1\sigma \ (2D), \ 3.2\sigma \ (3D) \tag{26}$$

provided that reaction rates are chosen according to [74]. For a proof that covers also reversible reactions, see [75]. Then the result is independent of the reaction rate k_r if in addition microscopic reversibility and a physically consistent dissociation rate are required. Above this limit, by treating the problem as a multiscale problem, it is possible to arrive at alternative reaction rates by accounting for the level of spatial resolution and by matching certain statistics of the microscopic model [39, 43, 74, 75]. These reaction rates become scale-dependent and hence a function of the mesh resolution h, in addition to the microscopic parameters, gives superior agreement with the SDLR compared to the classical constant (25). With a non-local extension of the RDME, agreement with the SDLR model can be obtained also below h^* [43, 72] with good accuracy on fine spatial scales, but at a steeply increasing computational cost as we approach a microscopic lattice method. In [82] a convergent non-local RDME-type model is constructed as a consistent discretization of the Doi model.

In summary, special care has to be taken to ensure that a lattice-based RDME model approaches one of the microscopic alternatives, and the expression (25) is not the best choice for strongly diffusion-limited processes. Fortunately, there are scale dependent alternatives that provide better results, but unfortunately there is to date no simple a priori way of determining what mesh resolution will be needed given some simulation accuracy tolerance and given a specific domain and set of reactions.

4 Macroscopic Models

When the copy numbers x_i are large in Sect. 3, a very good approximation of the concentrations of the chemical species $c_i(\mathbf{r}, t) = x_i/|\Omega|$, $i = 1, \ldots, M$, satisfies a reaction-diffusion equation in the domain Ω with boundary $\partial\Omega$

$$\frac{\partial \mathbf{c}}{\partial t} = \nabla \cdot (\boldsymbol{\gamma} \nabla \mathbf{c}) + \mathbf{f}(\mathbf{c}, \mathbf{r}), \ \mathbf{r} \in \Omega, \ t \geq 0. \tag{27}$$

The diffusion tensor $\boldsymbol{\gamma}$ is a diagonal matrix here which may depend on $\mathbf{r}$. In the simplest case, $\boldsymbol{\gamma} = \gamma \mathbf{I}$ and the diffusion is the same for all species everywhere in Ω. On $\partial\Omega$ with the normal $\mathbf{n}$, the molecules are reflected back into Ω with a Neumann condition

$$\mathbf{n} \cdot \nabla \mathbf{c} = 0, \tag{28}$$

and absorbed with a Dirichlet condition

$$\mathbf{c} = \mathbf{0}. \tag{29}$$

Combinations of these conditions are possible on $\partial\Omega$. The reaction term $\mathbf{f}$ in (27) depends on the propensities and the stoichiometry

$$\mathbf{f}(\mathbf{c}, \mathbf{r}) = \sum_{r=1}^{R} a_r(\mathbf{c}, \mathbf{r})(\boldsymbol{\nu}_r' - \boldsymbol{\nu}_r). \tag{30}$$

The concentration $\mathbf{c}$ is independent of $\mathbf{r}$ in a well-stirred system with a space independent $\mathbf{f}$. The resulting system of ODEs is then the reaction rate equations

$$\frac{d\mathbf{c}}{dt} = \mathbf{f}(\mathbf{c}), \ t \geq 0. \tag{31}$$

If $\mathbf{f}$ in (30) is an affine function of $\mathbf{c}$, then the mean values of the mesoscopic concentrations satisfy (27). With a nonlinear $\mathbf{f}$ due to, e.g., a bimolecular reaction, the solution of (27) is only an approximation of the mesoscopic mean values [33, 142].

The system (31) obeys mass-action kinetics if the propensities are

$$a_r(\mathbf{c}) = \kappa_r \prod_{m=1}^{M} c_m^{\nu_{rm}} \tag{32}$$

with a constant $\kappa_r > 0$ that in the case of diffusion-limited kinetics and a bimolecular reaction corresponds to the macroscopic rate constant k in (25). The mesoscopic coefficient is $\kappa_r/|\Omega|^{\sum_m \nu_{rm}-1}$, cf. k_a in (12) where $\sum_m \nu_{rm} = 2$. The steady state solution $\mathbf{c}_\infty = \lim_{t\to\infty} \mathbf{c}(t)$ of (31) with propensities as in (32) has interesting properties independent of κ_r if sufficient conditions on the reactions $\mathcal{R}$ in [3, 44] are satisfied. The graph of the reaction network must have a certain structure and a weak form of reversibility of the reactions is required. The steady state is then, e.g., unique and it is asymptotically stable for any choice of κ_r.

The mean square displacement (MSD) $\langle \|\mathbf{r}\|^2(t)\rangle$ of a molecule starting at $\mathbf{r} = 0$ at $t = 0$ and diffusing by Brownian dynamics is proportional to t. In observations of, e.g., crowded environments with fixed or mobile obstacles [8, 79, 104, 109, 116, 128], the MSD behaves anomalously like

$$\langle \|\mathbf{r}\|^2(t)\rangle \sim t^\alpha, \tag{33}$$

at least in a time window in a transient phase with $\alpha \in (0, 1)$. We have subdiffusion for $\alpha < 1$ with a lower MSD for large t than in ordinary diffusion with $\alpha = 1$. The corresponding macroscopic equation for subdiffusion is then a fractional PDE [109]

$$\partial_t^\alpha \mathbf{c} = \nabla \cdot (\boldsymbol{\gamma} \nabla \mathbf{c}), \ \mathbf{r} \in \Omega, \ t \geq 0. \tag{34}$$

The fractional time derivative is defined according to Caputo (C) or Riemann–Liouville (R-L)

$$\frac{\partial^\alpha u}{\partial t^\alpha} = \int_0^t \frac{(t-\tau)^{-\alpha}}{\Gamma(1-\alpha)} \partial_t u(\tau)\, d\tau \ \text{(C)}, \quad \frac{\partial^\alpha u}{\partial t^\alpha} = \partial_t \int_0^t \frac{(t-\tau)^{-\alpha}}{\Gamma(1-\alpha)} u(\tau)\, d\tau \ \text{(R-L)}. \tag{35}$$

Reactions with linear propensities can be introduced in this framework in different ways, see, e.g., [76]. A mesoscopic model with internal states is proposed in [111] and is extended to reactions in [12]. It does not seem to be clear how to include general reactions involving many species with nonlinear propensities.

The reaction-diffusion equation (27) is discretized in space by numerical solution with a finite element method (FEM) with linear elements [139], a finite volume method (FVM), or a finite difference method (FDM). The diffusion operator can be approximated by all three methods on a Cartesian mesh but if Ω is covered by a more flexible unstructured mesh, then FEM or FVM is the preferred choice.

Assume that the N components of $\mathbf{c}_m(t)$ are the nodal values of the concentrations of the chemical species m in the mesh. After spatial discretization, (27) will be

$$\partial_t \mathbf{c}_m = \mathbf{D}\mathbf{c}_m + \tilde{\mathbf{f}}_m(\mathbf{c}, \mathbf{r}), \ t \geq 0, \ m = 1, \ldots, M, \tag{36}$$

where $\tilde{\mathbf{f}}_m$ consists of the discretized reaction terms for species m. The elements of the diffusion matrix $\mathbf{D}$ are such that $D_{ij} \neq 0$, $i \neq j$, only if voxels $\mathcal{V}_i$ and $\mathcal{V}_j$ share a common edge (2D) or facet (3D) in the mesh, cf. q_{kj} in (21). If the boundary condition is (28), then the diagonal elements satisfy $D_{jj} = -\sum_{i=1}^{N} D_{ij} < 0$. Discretization by FDM on a Cartesian mesh with constant step size h yields the coefficients $D_{ij} = \gamma/h^2$ when $i \neq j$ and $D_{jj} = -2d\gamma/h^2$ where d is the dimension. Using FEM and FVM, $\mathbf{D}$ can be written as

$$\mathbf{D} = \mathbf{A}^{-1}\mathbf{S}, \tag{37}$$

where $\mathbf{A}$ is a diagonal matrix with $A_{ii} = |\mathcal{V}_i|$, the area or volume of $\mathcal{V}_i$. If FEM is used, $\mathbf{A}$ is a lumped mass matrix [139]. The stiffness matrix $\mathbf{S}$ depends on $\boldsymbol{\gamma}$ and the geometry of the discretization.

Based on the macroscopic discretization in (37), the mesoscopic jump coefficients between $\mathcal{V}_j$ and $\mathcal{V}_i$ in Sect. 3 are in [38] chosen to be

$$q_{ji} = D_{ij}\frac{|\mathcal{V}_i|}{|\mathcal{V}_j|} = \frac{S_{ij}}{|\mathcal{V}_j|}. \tag{38}$$

To be interpreted as jump probabilities, q_{ji} has to be non-negative. This is not always the case in a FEM discretization on a general unstructured mesh of poor quality where S_{ij} may be negative. The requirement that off-diagonal elements in $\mathbf{S}$ are non-negative is a sufficient condition for $\mathbf{c}_m$ in (36) with $\tilde{\mathbf{f}}_m = 0$ to satisfy the discrete maximum principle [139]. Then with a non-negative initial solution $c_{mi}(0) \geq 0$, $m = 1, \ldots, M$, $i = 1, \ldots, N$, the solution will stay non-negative $c_{mi}(t) \geq 0$ for all m and i when $t > 0$ and maxima in Ω will not increase and minima will not decrease. The same maximum principle is satisfied by the analytical solution of the corresponding parabolic PDE. How to modify the FEM discretization to achieve $S_{ij} \geq 0$ and to improve the mesh generation are discussed in, e.g., [13, 41, 94]. An alternative to deriving q_{ji} from (37) is to use the first exit time as in [102] resulting in non-negative q_{ji}. In [107] a slightly different problem with a modified diffusion tensor $\boldsymbol{\gamma}$ is solved, guaranteeing that $q_{ji} \geq 0$. Jump coefficients in a Cartesian mesh with a boundary cutting irregularly through the mesh are derived in [83] with the immersed boundary method.

The time scale of some of the reactions in (36) is sometimes much faster than the time scale of the other reactions and the diffusion. The ODE system in (36) is then stiff calling for dedicated numerical methods suitable for such problems. Another numerical difficulty is the nonlinearity in $\tilde{\mathbf{f}}_m$. By splitting the operator on the right-hand side and alternately solve for the diffusion and then solve for the reactions in a time step of length Δt the discretization error is proportional to Δt^2 in Strang splitting [132]. Each separate step is then computationally much simpler than a direct solution of the full set of time discretized equations. This procedure works well for moderately stiff problems but may cause a loss of accuracy for very stiff equations [131].

The relation between the solution of the ODE system in (36) and the mesoscopic model in Sect. 3 is investigated in [7, 95]. It is proved that if the mesoscopic system

approaches the thermodynamic limit and the system size grows, then its concentrations converge to the concentrations solving the reaction rate equations (31). In our case, the system is the ODE system in (36) for a given space discretization with voxels $\mathcal{V}_i,\ i = 1, \ldots, N$.

5 Hybrid and Multiscale Models

As a consequence of the inherent complexity of biochemical pathways in a cellular setting and the very disparate scales in reaction rates, diffusion constants, and molecule copy numbers, simulation of cellular control systems is a multiscale problem. Simplifications are possible in such systems, reducing the computational complexity. For example, two levels of modeling can be combined in a hybrid model incorporating two different scales of the mathematical representation of the system. In some cases, it may be necessary to include an accurate description of only a few molecular species at the microscopic level while most of the species can be well represented at the mesoscopic level, suggesting a micro–meso hybrid model. Large numbers of molecules of certain species are well modeled deterministically and should be simulated on the macroscale, but in the same model it might be necessary to resolve a few species where the randomness is important to capture. Then a hybrid model between the meso level and the macro level is recommended.

5.1 Microscopic–Mesoscopic Methods

By blending the microscopic and mesoscopic level, simulations can be constructed that attain high level of detail in critical areas of space or that treat strongly diffusion-limited kinetics accurately without resorting to a full Brownian Dynamics simulation or a very fine space discretization.

In [39, 43, 73], simulations are conducted on lattices, but with reaction propensities chosen systematically to match properties of the microscopic model. With appropriately chosen reaction rates, the mesoscopic model is transformed into a microscopic lattice model. These approaches have been discussed in Sect. 3.5.

In [73], an adaptive hybrid method is proposed that uses a background discretization of the domain with an unstructured mesh, and then partitions the system into either microscopic or mesoscopic degrees of freedom in each voxel of the mesh. The mesoscopic part is simulation with the RDME as implemented in URDME [26] and the microscopic part with GFRD [143] as implemented in [71]. The system can be split such that the domain is partitioned with interfaces between the microscopic and mesoscopic regimes, but also based on species so that strongly diffusion-limited reactions are captured on the microscopic level, avoiding the problems discussed in Sect. 3.5. In [92] the RDME is combined with Brownian Dynamics with bimolecular reactions implemented according to [100].

In [47, 48], the focus is on the diffusive interface, providing accurate transitions between the microscopic and the mesoscopic regimes. The method is extended to moving interfaces in [123]. The off-lattice software Smoldyn in Sect. 2.2 is enhanced in [124] by the hybrid method in [47, 123] for simulation of coupled microscopic and mesoscopic models.

5.2 Mesoscopic–Macroscopic Models

Mesoscopic simulation of a chemical reaction network with SSA may be very expensive if the copy numbers are large and if there is a considerable variation in the time scales. The different time scales in SSA simulations are due to the differences in reaction propensities. They depend on the reaction rate constants κ_r and the copy numbers of the species involved in (32).

In [67], ODEs are solved for the fast reactions and the slow reactions are modeled by a Markov process. There are fast numerical methods for stiff ODEs avoiding all the small time steps in the original SSA, thus saving computer time.

Another way of splitting the system is by treating species with large copy numbers as concentrations while species with small copy numbers such as genes as discrete variables. The motivation for this is the law of large numbers in [95].

A scaling parameter Q in the split system is introduced in [7, 87]. When $Q \to \infty$ convergence is proved to a limit system. The time scale of a species is determined by Q and the time scale of a reaction is derived from the scaling of the reaction rate and the copy numbers of the species involved in the reaction. The analysis in [87] is the basis for an algorithm in [78]. The reactions are split into continuously advanced reactions and a set of reactions firing at discrete events. The ODE system is integrated with discontinuous jumps in the right-hand side caused by the discrete events in a piecewise deterministic Markov process (PDMP) [20]. The species are measured either as continuous concentrations or as discrete copy numbers. The algorithm adaptively changes the scaling when the copy numbers change in the time interval of interest. A different but related idea is analyzed in [36], where the macroscopic model is employed in the form of a preconditioner to the mesoscopic one, thus bringing parallelism to an otherwise fully serial stochastic description.

The partitioning of the reactions into a fast set and a slow set is proposed also in a similar algorithm in [52, 113]. The error in the partitioning is estimated and determines when a reaction switches between the two sets. Another PDMP algorithm reducing the computing time is found in [18] where it is applied to gene networks. In [86], the rate of convergence is determined of marginal distributions computed by a PDMP. The rate is inversely proportional to a size parameter.

An alternative is to approximate the reaction network by a stochastic differential equation using Langevin dynamics for the fast reactions and SSA for the other reactions as in [127]. This approach is further developed with a blending region and is analyzed in [27].

Space dependence and diffusion introduce a special structure in the discretized mesoscopic model. Using operator splitting, reaction events are simulated with SSA and diffusion is treated with SSA, tau leaping, or macroscopic equations depending on estimates of the error in the approximations in [38, 46]. A path-wise stochastic analysis of similar split-step and multiscale approximations has very recently been developed [17, 37].

An analysis of a coupling scheme between the microscopic and the macroscopic levels of approximation is found in [49].

6 Discussion

We have reviewed computational methods for stochastic and deterministic models for biochemical reaction networks in cells. These models have different levels of accuracy, from the microscopic or off-lattice level, via the mesoscopic or on-lattice level to the macroscopic or PDE level of modeling. The focus is on systems with a spatial variation where molecules are transported by diffusion. Then Monte Carlo simulation is the only viable alternative for the *forward* problem to study the stochastic systems. A few areas of research related to the issues in this review are briefly mentioned below.

The models have parameters and initial conditions assumed to be known from experiments. The *inverse* problem of inferring the parameters given often noisy measurements has been addressed for the well-stirred problem, see, e.g., [19, 50, 93, 98, 99, 125, 126, 129, 148, 149], and for diffusion [120]. Parameter inference requires multiple solutions of the forward problem and comparison with data thus increasing the computational burden substantially.

If the diffusion parameter of a species is measured in vitro, then the observed diffusion in vivo is usually much slower. The reason is the excluded volume effect caused by other large molecules in the cell not included in the model [8, 79, 104]. Such molecular obstacles occupy up to 40% of the cellular volume. Crowding is introduced in the microscopic model in a natural way by including chemically inert molecules. At the mesoscopic level, the molecules are point particles and the diffusion rate has to be lowered somehow to account for the crowding. In [137] a volume-excluding compartment-based method that uses lattices on multiple resolutions is proposed. An open modeling question is how the reaction rates change at the mesoscopic and macroscopic levels.

The majority of studies and methods developed to date focus on diffusive transport, with some exceptions considering also active transport. In a spatial model based on PDEs, the role of active transport was discussed for the Hes1 and p53-MdM2 networks in [134]. Since most intracellular processes in eukaryotes involve a combination of diffusion and active motor-driven transport, more methodological research is needed to provide simulation methods capable of capturing also these effects. In [70] the RDME was extended to include active transport on fibers modeled through a mesoscopic velocity field and an advection term at the macro-

scopic level. A GPU implementation based on diffusion-drift stochastic differential equations was proposed in [145].

Another grand challenge for spatial stochastic simulations will be to couple the intracellular kinetics to biomechanics in order to model, e.g., growing domains or phenomena such as mechanosensing in which mechanical stimuli from the environment or touching cells provide input to the intracellular pathways [80].

Acknowledgements Generous support has been received from the Swedish Research Council, the eSSENCE strategic collaboration on e-Science, the UPMARC Linnaeus Center of Excellence, and the NIH under grant no. 1R01EB014877-01. The contents are solely the responsibility of the authors and do not necessarily reflect the opinions of these agencies. The authors would also like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Stochastic Dynamical Systems in Biology: Numerical Methods and Applications where this paper was conceived. This programme was supported by EPSRC grant no EP/K032208/1.

References

1. I.C. Agbanusi, S.A. Isaacson, A comparison of bimolecular reaction models for stochastic reaction diffusion systems. Bull. Math. Biol. **76**, 922–946 (2014)
2. D.F. Anderson, D.J. Higham, Multilevel Monte Carlo for continuous Markov chains, with applications in biochemical kinetics. Multiscale Model. Simul. **10**, 146–179 (2012)
3. D.F. Anderson, G. Craciun, T.G. Kurtz, Product-form stationary distributions for deficiency zero chemical reaction networks. Bull. Math. Biol. **72**, 1947–1970 (2010)
4. S.S. Andrews, D. Bray, Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. Phys. Biol. **1**, 137–151 (2004)
5. S.S. Andrews, N.J. Addy, R. Brent, A.P. Arkin, Detailed simulations of cell biology with Smoldyn 2.1. PLoS Comput. Biol. **6**(3), e1000705 (2010)
6. G. Arampatzis, M. Katsoulakis, P. Plecháč, Parallelization, processor communication and error analysis in lattice kinetic Monte Carlo. SIAM J. Numer. Anal. **52**(3), 1156–1182 (2014)
7. K. Ball, T.G. Kurtz, L. Popovic, G. Rempala, Asymptotic analysis of multiscale approximations to reaction networks. Ann. Appl. Probab. **16**, 1925–1961 (2006)
8. E. Barkai, Y. Garini, R. Metzler, Strange kinetics of single molecules in living cells. Phys. Today **65**(8), 29–35 (2012)
9. P. Bauer, J. Lindén, S. Engblom, B. Jonsson, Efficient inter-process synchronization for parallel discrete event simulation on multicores, in *Proceedings of the 3rd ACM SIGSIM Conference on Principles of Advanced Discrete Simulation, SIGSIM PADS '15*, pp. 183–194 (2015)
10. B. Bayati, P. Chatelin, P. Koumoutsakos, Adaptive mesh refinement for stochastic reaction-diffusion processes. J. Comput. Phys. **230**, 13–26 (2011)
11. U.S. Bhalla, Signaling in small subcellular volumes. I. Stochastic and diffusion effects on individual pathways. Biophys. J. **87**, 733–744 (2004)
12. E. Blanc, S. Engblom, A. Hellander, P. Lötstedt, Mesoscopic modeling of reaction-diffusion kinetics in the subdiffusive regime. Multiscale Model. Simul. **14**, 668–707 (2016)
13. J. Brandts, S. Korotov, M. Křížek, J. Šolc, On nonobtuse simplicial partitions. SIAM Rev. **51**(2), 317–335 (2009)
14. K. Burrage, J. Hancock, A. Leier, D.V. Nicolau Jr., Modelling and simulation techniques for membrane biology. Brief. Bioinform. **8**(4), 234–244 (2007)

15. Y. Cao, D.T. Gillespie, L.R. Petzold, The slow-scale stochastic simulation algorithm. J. Chem. Phys. **122**, 014116 (2005)
16. Y. Cao, D. Gillespie, L. Petzold, Efficient step size selection for the tau-leaping simulation method. J. Chem. Phys. **124**, 044109 (2006)
17. A. Chevallier, S. Engblom, Pathwise error bounds in multiscale variable splitting methods for spatial stochastic kinetics. Technical Report arXiv:1607.00805, Uppsala University, Uppsala (2016)
18. A. Crudu, A. Debussche, O. Radulescu, Hybrid stochastic simplifications for multiscale gene networks. BMC Syst. Biol. **3**, 89 (2009)
19. B.J. Daigle, M.K. Roh, L.R. Petzold, J. Niemi, Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. BMC Bioinf. **13**(1), 1–18 (2012)
20. M.H.A. Davies, Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models. J. R. Stat. Soc., Ser. B **46**, 358–388 (1984)
21. P. Deuflhard, W. Huisinga, T. Jahnke, M. Wulkow, Adaptive discrete Galerkin methods applied to the chemical master equation. SIAM J. Sci. Comput. **30**(6), 2990–3011 (2008)
22. M. Dobrzyński, J.V. Rodríguez, J.A. Kaandorp, J.G. Blom, Computational methods for diffusion-influenced biochemical reactions. Bioinformatics **23**(15), 134–155 (2007)
23. M. Doi, Stochastic theory of diffusion-controlled reaction. J. Phys. A: Math. Gen. **9**(9), 1479–1495 (1976)
24. A. Donev, V.V. Bulatov, T. Oppelstrup, G.H. Gilmer, B. Sadigh, M.H. Kalos, A First Passage Kinetic Monte Carlo algorithm for complex diffusion–reaction systems. J. Comput. Phys. **229**, 3214–3236 (2010)
25. B. Drawert, M.J. Lawson, L. Petzold, M. Khammash, The diffusive finite state projection algorithm for efficient simulation of the stochastic reaction-diffusion master equation. J. Chem. Phys. **132**(7), 074101 (2010)
26. B. Drawert, S. Engblom, A. Hellander, URDME: a modular framework for stochastic simulation of reaction-transport processes in complex geometries. BMC Syst. Biol. **6**, 76 (2012)
27. A. Duncan, R. Erban, K. Zygalakis, Hybrid framework for the simulation of stochastic chemical kinetics. J. Comput. Phys. **326**, 398–419 (2016)
28. W. E, D. Liu, E. Vanden-Eijnden, Nested stochastic simulation algorithm for chemical kinetic systems with disparate rates. J. Comput. Phys. **221**, 158–180 (2007)
29. J. Elf, M. Ehrenberg, Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. Genome Res. **13**, 2475–2484 (2003)
30. J. Elf, M. Ehrenberg, Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases. Syst. Biol. **1**, 230–236 (2004)
31. M.B. Elowitz, A.J. Levine, E.D. Siggia, P.S. Swain, Stochastic gene expression in a single cell. Science **297**, 1183–1186 (2002)
32. Y. Elskens, Microscopic derivation of a Markovian master equation in a deterministic model of chemical reaction. J. Stat. Phys. **37**(5–6), 673–695 (1984)
33. S. Engblom, Computing the moments of high dimensional solutions of the master equation. Appl. Math. Comput. **180**(2), 498–515 (2006)
34. S. Engblom, Numerical Solution Methods in Stochastic Chemical Kinetics. PhD thesis, Uppsala University (2008)
35. S. Engblom, Spectral approximation of solutions to the chemical master equation. J. Comput. Appl. Math. **229**(1), 208–221 (2009)
36. S. Engblom, Parallel in time simulation of multiscale stochastic chemical kinetics. Multiscale Model. Simul. **8**(1), 46–68 (2009)
37. S. Engblom, Strong convergence for split-step methods in stochastic jump kinetics. SIAM J. Numer. Anal. **53**(6), 2655–2676 (2015)
38. S. Engblom, L. Ferm, A. Hellander, P. Lötstedt, Simulation of stochastic reaction-diffusion processes on unstructured meshes. SIAM J. Sci. Comput. **31**, 1774–1797 (2009)
39. R. Erban, J. Chapman, Stochastic modelling of reaction-diffusion processes: algorithms for bimolecular reactions. Phys. Biol. **6**, 046001 (2009)

40. R. Erban, J. Chapman, P.K. Maini, A practical guide to stochastic simulations of reaction-diffusion processes. Technical report, Mathematical Institute, University of Oxford, Oxford (2007). http://arxiv.org/abs/0704.1908
41. H. Erten, A. Üngör, Quality triangulations with locally optimal Steiner points. SIAM J. Sci. Comput. **31**, 2103–2130 (2009)
42. D. Fange, J. Elf, Noise-induced min phenotypes in *E. coli*. PLoS Comput. Biol. **2**, 637–648 (2006)
43. D. Fange, O.G. Berg, P. Sjöberg, J. Elf, Stochastic reaction-diffusion kinetics in the microscopic limit. Proc. Natl. Acad. Sci. U. S. A. **107**(46), 19820–19825 (2010)
44. M. Feinberg, The existence and uniqueness of steady states for a class of chemical reaction networks. Arch. Ration. Mech. Anal. **132**, 311–370 (1995)
45. L. Ferm, P. Lötstedt, A. Hellander, A hierarchy of approximations of the master equation scaled by a size parameter. J. Sci. Comput. **34**, 127–151 (2008)
46. L. Ferm, A. Hellander, P. Lötstedt, An adaptive algorithm for simulation of stochastic reaction-diffusion processes. J. Comput. Phys. **229**, 343–360 (2010)
47. M.B. Flegg, S.J. Chapman, R. Erban, The two-regime method for optimizing stochastic reaction-diffusion simulations. J. R. Soc. Interface **9**, 859–868 (2012)
48. M.B. Flegg, S. Hellander, R. Erban, Convergence of methods for coupling of microscopic and mesoscopic reaction–diffusion simulations. J. Comput. Phys. **289**, 1–17 (2015)
49. B. Franz, M. Flegg, S.J. Chapman, R. Erban, Multiscale reaction-diffusion algorithms: PDE-assisted Brownian dynamics. SIAM J. Appl. Math. **73**, 1224–1247 (2013)
50. F. Fröhlich, P. Thomas, A. Kazeroonian, F.J. Thies, R. Grima, J. Hasenauer, Inference for stochastic chemical kinetics using moment equations and systems size expansion. PLoS Comput. Biol. **12**, e1005030 (2016)
51. C. Gadgil, C.-H.- Lee, H.G. Othmer, A stochastic analysis of first-order reaction networks. Bull. Math. Biol. **67**, 901–946 (2005)
52. A. Ganguly, D. Altintan, H. Koeppl, Jump-diffusion approximation of stochastic reaction dynamics: Error bounds and algorithms. Multiscale Model. Simul. **13**, 1390–1419 (2015)
53. C.W. Gardiner, *Handbook of Stochastic Methods*. Springer Series in Synergetics, 3rd edn. (Springer, Berlin 2004)
54. C.W. Gardiner, K.J. McNeil, D.F. Walls, I.S. Matheson, Correlations in stochastic theories of chemical reactions. J. Stat. Phys. **14**(4), 307–331 (1976)
55. M.A. Gibson, J. Bruck, Efficient exact stochastic simulation of chemical systems with many species and many channels. J. Phys. Chem. **104**(9), 1876–1889 (2000)
56. D.T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. **22**(4), 403–434 (1976)
57. D.T. Gillespie, A rigorous derivation of the chemical master equation. Physica A **188**, 404–425 (1992)
58. D.T. Gillespie, *Markov Processes: An introduction for Physical Scientists* (Academic, San Diego, CA, 1992)
59. D.T. Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems. J. Chem. Phys. **115**(4), 1716–1733 (2001)
60. D.T. Gillespie, A diffusional bimolecular propensity function. J. Chem. Phys. **131**(16), 164109 (2009)
61. D.T. Gillespie, E. Seitaridou, *Simple Brownian Diffusion* (Oxford University Press, Oxford, 2013)
62. D.T. Gillespie, A. Hellander, L.R. Petzold, Perspective: stochastic algorithms for chemical kinetics. J. Chem. Phys. **138**, 170901 (2013)
63. J. Goutsias, G. Jenkinson, Markovian dynamics on complex reaction networks. Phys. Rep. **529**, 199–264 (2013)
64. R. Grima, A study of the moment-closure approximations for stochastic chemical kinetics. J. Chem. Phys. **136**, 154105 (2012)
65. R. Grima, Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. Phys. Rev. E **92**, 042124 (2015)

66. P. Hammar, P. Leroy, A. Mahmutovic, E.G. Marklund, O.G. Berg, J. Elf, The *lac* repressor displays facilitated diffusion in living cells. Science **336**, 1595–1598 (2012)
67. E.L. Haseltine, J.B. Rawlings, Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. J. Chem. Phys. **117**(15), 6959–6969 (2002)
68. J. Hattne, D. Fange, J. Elf, Stochastic reaction-diffusion simulation with MesoRD. Bioinformatics **21**, 2923–2924 (2005)
69. M. Hegland, C. Burden, L. Santoso, S. MacNamara, H. Booth, A solver for the stochastic master equation applied to gene regulatory networks. J. Comput. Appl. Math. **205**(2), 708–724 (2007)
70. A. Hellander, P. Lötstedt, Incorporating active transport of cellular cargo in stochastic mesoscopic models of living cells. Multiscale Model. Simul. **8**(5), 1691–1714 (2010)
71. S. Hellander, P. Lötstedt, Flexible single molecule simulation of reaction-diffusion processes. J. Comput. Phys. **230**, 3948–3965 (2011)
72. S. Hellander, L. Petzold, Reaction rates for a generalized reaction-diffusion master equation. Phys. Rev. E **93**, 013307 (2016)
73. A. Hellander, S. Hellander, P. Lötstedt, Coupled mesoscopic and microscopic simulation of stochastic reaction-diffusion processes in mixed dimensions. Multiscale Model. Simul. **10**(2), 585–611 (2012)
74. S. Hellander, A. Hellander, L. Petzold, Reaction-diffusion master equation in the microscopic limit. Phys. Rev. E **85**, 042901 (2012)
75. S. Hellander, A. Hellander, L. Petzold, Reaction rates for mesoscopic reaction-diffusion kinetics. Phys. Rev. E **91**, 023312 (2015)
76. B.J. Henry, T.A.M. Langlands, S.L. Wearne, Anomalous diffusion with linear reaction dynamics: from continuous time random walks to fractional reaction-diffusion equations. Phys. Rev. E **74**, 031116 (2006)
77. I. Hepburn, W. Chen, S. Wils, E.D. Schutter, STEPS: efficient simulation of stochastic reaction-diffusion models in realistic morphologies. BMC Syst. Biol. **6**, 36 (2012)
78. B. Hepp, A. Gupta, M. Khammash, Adaptive hybrid simulations for multiscale stochastic reaction networks. J. Chem. Phys. **142**(3), 034118 (2015)
79. F. Höfling, T. Franosch, Anomalous transport in the crowded world of biological cells. Rep. Progr. Phys. **76**, 046602 (2013)
80. J.D. Humphrey, E.R. Dufresne, M.A. Schwartz, Mechanotransduction and extracellular matrix homeostasis. Nat. Rev. Mol. Cell Biol. **15**(12), 802–812 (2014)
81. S.A. Isaacson, The reaction-diffusion master equation as an asymptotic approximation of diffusion to a small target. SIAM J. Appl. Math. **70**, 77–111 (2009)
82. S.A. Isaacson, A convergent reaction-diffusion master equation. J. Chem. Phys. **139**, 054101 (2013)
83. S.A. Isaacson, C.S. Peskin, Incorporating diffusion in complex geometries into stochastic chemical kinetics simulations. SIAM J. Sci. Comput. **28**(1), 47–74 (2006)
84. T. Jahnke, An adaptive wavelet method for the chemical master equation. SIAM J. Sci. Comput. **31**(6), 4373–4394 (2010)
85. T. Jahnke, W. Huisinga, Solving the chemical master equation for monomolecular reaction systems analytically. J. Math. Biol. **54**(1), 1–26 (2007)
86. T. Jahnke, M. Krein, Error bound for piecewise deterministic processes modeling stochastic reaction systems. Multiscale Model. Simul. **10**, 1119–1147 (2012)
87. H.-W. Kang, T.G. Kurtz, Separation of time-scales and model reduction for stochastic reaction networks. Ann. Appl. Probab. **23**, 529–583 (2013)
88. J.R. Karr, J.C. Sanghvi, D.N. Macklin, M.V. Gutschow, J.M. Jacobs, B. Bolival Jr., N. Assad-Garcia, J.I. Glass, M.W. Covert, A whole-cell computational model predicts phenotype from genotype. Cell **150**, 389–401 (2012)
89. V. Kazeev, M. Khammash, M. Nip, C. Schwab, Direct solution of the chemical master equation using quantized tensor trains. PLoS Comput. Biol. **10**(4), e1003359 (2014)
90. R.A. Kerr, T.M. Bartol, B. Kaminsky, M. Dittrich, J.-C.J. Chang, S.B. Baden, T.J. Sejnowski, J.R. Stiles, Fast Monte Carlo simulation methods for biological reaction-diffusion systems in

solution and on surfaces. SIAM J. Sci. Comput. **30**(6), 3126–3149 (2008)
91. D.J. Kiviet, P. Nghe, N. Walker, S. Boulineau, V. Sunderlikova, S.J. Tans, Stochasticity of metabolism and growth at the single-cell level. Nature **514**, 376–379 (2014)
92. M. Klann, A. Ganguly, H. Koeppl, Hybrid spatial Gillespie and particle tracking simulation. Bioinformatics **28**, i549–i555 (2012)
93. M. Komorowski, M.J. Costa, D.A. Rand, M.P.H. Stumpf, Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. Proc. Natl. Acad. Sci. U. S. A. **108**(21), 8645–8650 (2011)
94. S. Korotov, M. Křížek, P. Neittanmäki, Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle. Math. Comp. **70**, 107–119 (2000)
95. T.G. Kurtz, Solutions of ordinary differential equations as limits of pure jump Markov processes. J. Appl. Prob. **7**, 49–58 (1970)
96. T.G. Kurtz, Strong approximation theorems for density dependent Markov chains. Stoch. Proc. Appl. **6**, 223–240 (1978)
97. C. Lemerle, B.D. Ventura, L. Serrano, Space as the final frontier in stochastic simulations of biological systems. FEBS Lett. **579**, 1789–1794 (2005)
98. S. Liao, T. Vejchodsky, R. Erban, Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks. J. R. Soc. Interface **12**, 20150233 (2015)
99. G. Lillacci, M. Khammash, The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. Bioinformatics **29**(18), 2311–2319 (2013)
100. J. Lipková, K.C. Zygalakis, S.J. Chapman, R. Erban, Analysis of Brownian dynamics simulations of reversible bimolecular reactions. SIAM J. Appl. Math. **71**(3), 714–730 (2011)
101. C.F. Lopez, J.L. Muhlich, J.A. Bachman, P.K. Sorger, Programming biological models in python using PySB. Mol. Syst. Biol. **9**(1), 646 (2013)
102. P. Lötstedt, L. Meinecke, Simulation of stochastic diffusion via first exit times. J. Comput. Phys. **300**, 862–886 (2015)
103. A. Mahmutovic, D. Fange, O.G. Berg, J. Elf, Lost in presumption: stochastic reactions in spatial models. Nat. Methods **9**(12), 1–4 (2012)
104. T.T. Marquez-Lago, A. Leier, K. Burrage, Anomalous diffusion and multifractional Brownian motion: simulating molecular crowding and physical obstacles in systems biology. IET Syst. Biol. **6**(4), 134–142 (2012)
105. H.H. McAdams, A. Arkin, Stochastic mechanisms in gene expression. Proc. Natl. Acad. Sci. U. S. A. **94**, 814–819 (1997)
106. D.A. McQuarrie, Stochastic approach to chemical kinetics. J. Appl. Prob. **4**, 413–478 (1967)
107. L. Meinecke, S. Engblom, A. Hellander, P. Lötstedt, Analysis and design of jump coefficients in discrete stochastic diffusion models. SIAM J. Sci. Comput. **38**, A55–A83 (2016)
108. R. Metzler, The future is noisy: the role of spatial fluctuations in genetic switching. Phys. Rev. Lett. **87**, 068103 (2001)
109. R. Metzler, J. Klafter, The random walk's guide to anomalous diffusion: a fractional dynamics approach. Phys. Rep. **339**(1), 1–77 (2000)
110. A. Milias-Argeitis, S. Engblom, P. Bauer, M. Khammash, Stochastic focusing coupled with negative feedback enables robust regulation in biochemical reaction networks. J. R. Soc. Interface **12**(113), 1–10 (2015)
111. M.S. Mommer, D. Lebiedz, Modeling subdiffusion using reaction diffusion systems. SIAM J. Appl. Math. **70**(1), 112–132 (2009)
112. A. Moraes, R. Tempone, P. Vilanova, Hybrid Chernoff tau-leap. Multiscale Model. Simul. **12**(2), 581–615 (2014)
113. A. Moraes, R. Tempone, P. Vilanova, A multilevel adaptive reaction-splitting simulation method for stochastic reaction networks. SIAM J. Sci. Comput. **38**(4), A2091–A2117 (2016)
114. B. Munsky, M. Khammash, The finite state projection algorithm for the solution of the chemical master equation. J. Chem. Phys. **124**(4), 044104 (2006)
115. B. Munsky, G. Neuert, A. van Oudenaarden, Using gene expression noise to understand gene regulation. Science **336**(6078), 183–187 (2012)

116. D.V. Nicolau Jr., J.F. Hancock, K. Burrage, Sources of anomalous diffusion on cell membranes: a Monte Carlo study. Biophys. J. **92**, 1975–1987 (2007)
117. T. Oppelstrup, V.V. Bulatov, A. Donev, M.H. Kalos, G.H. Gilmer, B. Sadigh, First-passage kinetic Monte Carlo method. Phys. Rev. E **80**, 066701 (2009)
118. J. Pahle, Biochemical simulations: stochastic, approximate stochastic and hybrid approaches. Brief. Bioinform. **10**, 53–64 (2009)
119. J. Paulsson, Summing up the noise in gene networks. Nature **427**, 415–418 (2004)
120. F. Persson, M. Lindén, C. Unosson, J. Elf, Extracting intracellular diffusive states and transition rates from single-molecule tracking data. Nat. Methods **10**, 265–269 (2013)
121. A. Raj, A. van Oudenaarden, Nature, nurture, or chance: Stochastic gene expression and its consequences. Cell **135**(2), 216–226 (2008)
122. M. Rathinam, L.R. Petzold, Y. Cao, D. Gillespie, Consistency and stability of tau-leaping schemes for chemical reaction systems. Multiscale Model. Simul. **4**, 867–895 (2005)
123. M. Robinson, M. Flegg, R. Erban, Adaptive two-regime method: application to front propagation. J. Chem. Phys. **140**, 124109 (2014)
124. M. Robinson, S.S. Andrews, R. Erban, Multiscale reaction-diffusion simulations with Smoldyn. Bioinformatics **31**(14), 2406–2408 (2015)
125. J. Ruess, A. Milias-Argeitis, J. Lygeros, Designing experiments to understand the variability in biochemical reaction networks. J. R. Soc. Interface **10**(88) (2013)
126. J. Ruess, F. Parise, A. Milias-Argeitis, M. Khammash, J. Lygeros, Iterative experiment design guides the characterization of a light-inducible gene expression circuit. Proc. Natl. Acad. Sci. U. S. A. **112**(26), 8148–8153 (2015)
127. H. Salis, Y. Kaznessis, Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. J. Chem. Phys. **122**, 054103 (2005)
128. M.J. Saxton, A biological interpretation of transient anomalous subdiffusion. I. Qualitative model. Biophys. J. **92**, 1178–1191 (2007)
129. D. Schnoerr, G. Sanguinetti, R. Grima, Approximation and inference methods for stochastic biochemical kinetics - a tutorial review. Technical report, University of Edinburgh, Edinburgh (2016). http://arxiv.org/abs/1608.06582
130. J. Schöneberg, A. Ullrich, F. Noé, Simulation tools for particle-based reaction-diffusion dynamics in continuous space. BMC Biophysics **7**, 11 (2014)
131. B. Sportisse, An analysis of operator splitting techniques in the stiff case. J. Comput. Phys. **161**, 140–168 (2000)
132. G. Strang, On the construction and comparison of difference schemes. SIAM J. Numer. Anal. **5**, 506–517 (1968)
133. A.B. Stundzia, C.L. Lumsden, Stochastic simulation of coupled reaction-diffusion processes. J. Comput. Phys. **127**, 196–207 (1996)
134. M. Sturrock, A.J. Terry, D.P. Xirodimas, A.M. Thompson, M.A. Chaplain, Influence of the nuclear membrane, active transport, and cell shape on the Hes1 and p53-Mdm2 pathways: insights from spatio-temporal modelling. Bull. Math. Biol. **74**, 1531–1579 (2012)
135. K. Takahashi, S.N.V. Arjunan, M. Tomita, Space in systems biology of signaling pathways–towards intracellular molecular crowding in silico. FEBS Lett. **579**, 1782–1788 (2005)
136. K. Takahashi, S. Tănase-Nicola, P.R. ten Wolde, Spatio-temporal correlations can drastically change the response of a MAPK pathway. Proc. Natl. Acad. Sci. U. S. A. **107**(6), 2473–2478 (2010)
137. P.R. Taylor, R.E. Baker, M.J. Simpson, C.A. Yates, Coupling volume-excluding compartment-based models of diffusion at different scales: voronoi and pseudo-compartment approaches. J. R. Soc. Interface **13**(120), 20160336 (2016)
138. M. Thattai, A. van Oudenaarden, Intrinsic noise in gene regulatory networks. Proc. Nat. Acad. Sci. U. S. A. **98**, 8614–8619 (2001)
139. V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems* (Springer, Berlin, 1997)
140. T. Tian, K. Burrage, Binomial leap methods for simulating stochastic chemical kinetics. J. Chem. Phys. **121**, 10356–10364 (2004)

141. M. Tomita. Whole-cell simulation: a grand challenge for the 21st century. Trends Biotechnol. **19**(6), 205–210 (2001)
142. N.G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 5th edn. (Elsevier, Amsterdam, 2004)
143. J.S. van Zon, P.R. ten Wolde, Green's-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space. J. Chem. Phys. **123**, 234910 (2005)
144. M. von Smoluchowski, Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen. Z. Phys. Chem. **92**, 129–168 (1917)
145. M. Vigelius, B. Meyer, Multi-dimensional, mesoscopic Monte Carlo simulations of inhomogeneous reaction-drift-diffusion systems on graphics-processing units. PLoS One **7**(4), 1–13 (2012)
146. J.M.G. Vilar, H.Y. Kueh, N. Barkai, S. Leibler, Mechanism of noise-resistance in genetic oscillators. Proc. Nat. Acad. Sci. **99**, 5988–5992 (2002)
147. M. Watabe, S.N.V. Arjunan, S. Fukushima, K. Iwamoto, J. Kozuka, S. Matsuoka, Y. Shindo, M. Ueda, K. Takahashi, A computational framework for bioimaging simulation. PLoS One **10**(7), 1–19 (2015)
148. C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, H. Koeppl, Moment-based inference predicts bimodality in transient gene expression. Proc. Natl. Acad. Sci. U. S. A. **109**(21), 8340–8345 (2012)
149. C. Zechner, M. Unger, S. Pelet, M. Peter, H. Koeppl, Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. Nat. Methods **11**, 197–202 (2014)

Part II
Stochastic Numerical Approaches, Algorithms and Coarse-Grained Simulations

Numerical Methods for Stochastic Simulation: When Stochastic Integration Meets Geometric Numerical Integration

Assyr Abdulle

1 Introduction

In this paper we review a recently developed framework to construct and analyze efficient numerical methods to approximate expectation of a functional of stochastic processes. This is a basic problem for many applications in biology, chemistry or physics [15]. For example, in molecular dynamics where a fundamental issue is the computation of macroscopic quantities, typically functionals of some variables of the system with respect to a given probability measure often given by the Boltzmann-Gibbs density. The associated numerical problem consists in solving high-dimensional integrals that are most often approximated through ergodic averages of stochastic dynamics obtained from solutions of stochastic differential equations (SDEs), e.g., Langevin SDEs [24]. The approximation of functionals of a stochastic process also arise in multiscale stochastic systems. In the SDE context, one would like to solve, for example, systems of the type[1]

$$\begin{aligned} dX &= f(X,Y)dt, \\ dY &= \frac{1}{\varepsilon}g(X,Y)dt + \frac{1}{\sqrt{\varepsilon}}\sigma(X,Y)dW(t), \end{aligned}$$

where $W(t)$ is a Wiener process and ε is a parameter $\varepsilon \ll 1$. Classical stochastic solvers will need a time-step that resolves the fast dynamics resulting in a large

[1] See Sect. 1.1 for a precise definition.

A. Abdulle (✉)
ANMC, Institut de Mathmatiques, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
e-mail: assyr.abdulle@epfl.ch

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_4

computational cost. For classes of such multi-scale problems, averaging or homogenization techniques [33] can lead to more efficient numerical integrators. This is the case, for example, when an effective macro dynamics exists, i.e., when the existence of an equation of the form $d\bar{X} = \lim_{\varepsilon\to 0} \int f(\bar{X}, Y) d\mu^{\varepsilon}_{\bar{X}}(dY)dt = F(\bar{X})dt$, where $\mu^{\varepsilon}_{\bar{X}}$ is an invariant measure for the fast dynamics of the above system with $X = \bar{X}$ fixed can be established. In this situation, one can implement a multiscale scheme that consists in sampling the fast variable while the slow variable X_n is fixed at time t_n, perform an averaging to recover an approximation $\tilde{F}(X_n)$ of the force $F(X_n)$ and use a macroscopic solver to advance the effective dynamics to time t_{n+1}, e.g., $X_{n+1} = X_n + h\tilde{F}(X_n)$ (see [3, 13, 39]). Such multiscale methods have also been developed for stochastic partial differential equations [1, 8]. One of the main issues for such fast–slow numerical techniques is the computational cost of the repeated computation of the effective forces, relying on the approximation of the invariant measure of the fast process. Accurate approximation of invariant measures is a central issue in such stochastic computations [13, 26]. We note that an algorithm based on the solution of a Poisson problem obtained from the generator of the fast system is also promising for classes of SDEs (or SPDEs) with multiple scales [6]. For the numerical approximation of ergodic SDEs, one faces several important questions:

1. does the numerical method have an invariant measure ?
2. how close is the numerical invariant measure to the true one ?
3. how close is the time-averaging method to the invariant measure ?

In this paper we will mainly discuss the second question. We note that many authors have discussed the first question (see [27, 31, 32, 35, 38] and the references in these papers), for the third question there are also a large body of contributions and we refer to the textbook [17, 20, 30] for references.

Fast–slow processes are also ubiquitous in biology when simulating N chemical species $\{S_i\}_{i=1}^N$ interacting through M reaction channels $\{R_i\}_{i=1}^M$. The state of the system is specified by the vector $\mathbf{X}_t = (X_{1t}, \ldots, X_{Nt})^T$ that can be shown to evolve according to a Markov process. Sampling trajectories can be computed according to the stochastic simulation algorithm by updating the waiting time to the next reaction and selecting the next reaction that occurs. This numerical algorithm is called the stochastic simulation algorithm (or Gillespie algorithm, see [16] for a review). In a multiscale context, when some reaction channels occur frequently on a timescale for which others will only rarely take place, a large computational effort will be needed to see the dynamics of some slower reaction channels. In this context, multiscale algorithms that share similarities with the above SDE situation have been developed. We mention the nested SSA [14], the slow-scale SSA [10] that are both based on quasi-steady approximation (the time scale separation between fast and slow processes allows for the fast process to equilibrate before significant change in the slow process occur). Here again, part of the computational effort is devoted to compute the equilibrium of fast process.

The framework presented below for constructing and analyzing stochastic integrators for the computation of expectation of functionals of stochastic processes for both finite time or long-time dynamics has been introduced in [2, 4, 5]. While this framework has been applied to SDEs, applications to discrete stochastic processes might be an interesting topic to explore in the future.

1.1 Setting and Definitions

We consider a d-dimensional SDE

$$dX = f(X)dt + g(X)dW(t), \quad X(0) = X_0, \tag{1}$$

where $X_0 \in E$ is the initial condition assumed deterministic for simplicity, and $W(t)$ is a standard m-dimensional Wiener process. The maps $f : E \mapsto E$, $g : E \mapsto E^m$ are assumed to be smooth and the space E denotes either $E = \mathbb{R}^d$ or the torus $E = \mathbb{T}^d$, and is specified when needed. We will sometimes use the vector notation $g = (g^1, \ldots, g^m)$ for the matrix $g(x)$, where $g^i(x) \in E$.

We consider a discrete numerical approximation of (1) given by

$$X_{n+1} = \Psi(X_n, h, \xi_n), \tag{2}$$

for $X_n \in E$ for $n \geq 0$, where $\Psi(\cdot, h, \xi_n) : E \to E$ is the discrete numerical flow, h denotes the time-step size, and ξ_n denotes a random vector.

Several concepts of convergence can be used to measure how well the numerical method (2) approximates the solution of (1). We briefly review here strong and weak convergence, as well as convergence with respect to an invariant measure of (1) provided such invariant measures exist.

1.2 Strong and Weak Convergence

We note by $C_P^\ell(\mathbb{R}^d, \mathbb{R})$ the space of $1 \leq \ell \leq \infty$ times continuously differentiable functions $\mathbb{R}^d \to \mathbb{R}$ with all partial derivatives with polynomial growth. Likewise $C^\ell(\mathbb{T}^d, \mathbb{R})$ will denote the space of $1 \leq \ell \leq \infty$ times continuously differentiable functions $\mathbb{T}^d \to \mathbb{R}$. To simplify the notation, we will define $\mathcal{V}^\ell(E, \mathbb{R})$ to denote either $C_P^\ell(\mathbb{R}^d, \mathbb{R})$ or $C^\ell(\mathbb{T}^d, \mathbb{R})$ and $\mathcal{V}^\ell(E, E)$ to denote either $C_P^\ell(\mathbb{R}^d, \mathbb{R}^d)$ or $C^\ell(\mathbb{T}^d, \mathbb{T}^d)$. When needed the specific situation will be mentioned.

The numerical approximation (2), starting from the exact initial condition X_0 of (1) is said to have weak order p if for all functions $\phi \in \mathcal{V}^{2(p+1)}(E, \mathbb{R})$

$$|\mathbb{E}(\phi(X_n)) - \mathbb{E}(\phi(X(t_n)))| \leq Ch^p, \tag{3}$$

and to have strong order p if

$$\mathbb{E}(|X_n - X(t_n)|) \leq Ch^p, \tag{4}$$

for any $t_n = nh \in [0, T]$ with $T > 0$ fixed, for all h small enough, with constants C independent of h.

Remark 1 We will sometimes use the result [29] (see [30, Chap. 2.2]) of Milstein that allows to deduce global weak order from the local weak error, i.e., the weak error after one step. The results are as follows: assuming that $f, g^r \in \mathcal{V}^{2(p+1)}(E, E), r = 1, \ldots, m$ are Lipschitz continuous, that for all $r \in \mathbb{N}$, the moments $\mathbb{E}(|X_n|^{2r})$ are bounded for all n, h with $0 \le nh \le T$ uniformly with respect to all h sufficiently small (this condition only applies for the case $E = \mathbb{R}^d$), and that for all $\phi \in \mathcal{V}^{2(p+1)}(E, \mathbb{R})$ and all initial values $X(0) = X_0$ the local weak error satisfies

$$|\mathbb{E}(\phi(X_1)) - \mathbb{E}(\phi(X(t_1)))| \le Ch^{p+1} \tag{5}$$

for all h sufficiently small, then the global error bound (3) holds. For the strong error, conditions on local weak and strong errors are needed to infer global strong convergence (see [30] for details).

1.3 Long-Time Behavior: Approximation of the Invariant Measure

The above strong and weak convergence measure finite time approximation properties of the numerical solver (2) when applied to (1). Of interest in many applications is the long-time approximation of ergodic SDEs (1), i.e., SDEs that have a unique invariant measure μ satisfying for each smooth integrable function ϕ and for any deterministic initial condition X_0,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \phi(X(s))ds = \int_E \phi(y)d\mu(y), \qquad \text{almost surely.} \tag{6}$$

In this paper we will assume that the numerical method (2) is ergodic, i.e., that it has a unique invariant probability law μ^h with finite moments of any order and

$$\lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} \phi(X_n) = \int_E \phi(y)d\mu^h(y), \qquad \text{almost surely,} \tag{7}$$

for all deterministic initial condition X_0 and all smooth test functions ϕ. To quantify the second question we will say that the numerical method (2) has order $p \ge 1$ with respect to the invariant measure if

$$|e(\phi, h)| \le Ch^p \quad \text{with} \quad e(\phi, h) := \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} \phi(X_n) - \int_E \phi(y)d\mu(y), \tag{8}$$

where h is small enough and C is independent of h and X_0 and the above limit holds with probability one.

2 Tools Inspired from Geometric Numerical Integration

In this section we recall some classical tools of geometric numerical integration for ODEs, developed in particular for the analysis of the long-time dynamics of symplectic integrators. We refer to [18] for a comprehensive presentation and detailed proofs of the claims made below. Consider an ordinary differential equation written in autonomous form for simplicity

$$\begin{aligned} \frac{dX}{dt} &= f(X) \\ X(0) &= X_0, \end{aligned} \tag{9}$$

where $t \geq 0, f : E \longrightarrow E$. We denote by $\varphi_t(x)$ the exact flow of this differential equation. We consider a one-step numerical method with constant time-step h

$$X_{n+1} = \psi_h(X_n; f), \tag{10}$$

where $\psi_h(\cdot; f) : E \longrightarrow E$ is the discrete numerical flow and the second argument of ψ_h just emphasizes the differential equation the numerical integrator is applied to. We recall that the numerical method (10) is said to have order p if

$$\psi_h(x; f) - \varphi_h(x) = \mathcal{O}(h^{p+1}), \tag{11}$$

for all sufficiently smooth differential equations (9).

2.1 Backward Error Analysis

We assume that the numerical method can be expanded as

$$\psi_h(x; f) = x + hf(x) + h^2 d_1(x) + h^3 d_2(x) + \cdots \tag{12}$$

We note that the form of the two first terms in the above expansion is required for a consistent method. The idea of backward error analysis is to find a modified differential equation

$$\frac{d\widetilde{X}}{dt} = \tilde{f}_{h,s}(\widetilde{X}),$$

$$\widetilde{X}(0) = X_0, \tag{13}$$

where[2]

$$\tilde{f}_{h,s}(x) = f(x) + h\tilde{f}_1(x) + h^2\tilde{f}_2(x) + \cdots + h^s\tilde{f}_s(x), \tag{14}$$

and to construct $\tilde{f}_i(x)$ such that $\widetilde{X}(h) - \psi_h(x;f) = \mathcal{O}(h^{s+2})$ for $h \to 0$. It then follows that $\widetilde{X}(t_n) = X_n + \mathcal{O}(h^{s+1})$ for $h \to 0$ and bounded times $t_n = nh \le T$. We denote by $\tilde{\varphi}_t(x)$ the flow of the differential equation (13). To compute the function $\tilde{f}_2, \tilde{f}_3, \ldots$ one expands the solution of (13) into a Taylor series and compare the power of h with (12). In this respect, we have the following lemma

Lemma 1 *If the numerical method (10) is of order p with $\psi_h(x;f) - \varphi_h(x) = h^{p+1}\delta_{p+1}(x) + \mathcal{O}(h^{p+2})$, then (assuming $s \ge p$) the function $\tilde{f}_{h,s}$ given by (14) satisfies*

$$\tilde{f}_{h,s}(x) = f(x) + h^p\tilde{f}_p(x) + \cdots h^s\tilde{f}_s(x), \tag{15}$$

where $\tilde{f}_p(x) = \delta_{p+1}(x)$.

We next rewrite the exact and modified flows in terms of differential operators. If we define the Lie derivative by $\mathcal{L}_D = f \cdot \nabla$ then, provided that f is $M+1$ times continuously differentiable, the flow of (9) satisfies

$$\phi(\varphi_h(x)) = \sum_{k=0}^{M} \frac{h^k}{k!}(\mathcal{L}_D^k\phi)(x) + \mathcal{O}(h^{M+1}), \tag{16}$$

for all smooth test functions ϕ. Likewise for the modified Eq. (13) we have

$$\phi(\tilde{\varphi}_h(x)) = \sum_{k=0}^{M} \frac{h^k}{k!}(\widetilde{\mathcal{L}}_D^k\phi)(x) + \mathcal{O}(h^{M+1}), \tag{17}$$

where $\widetilde{\mathcal{L}}_D = \tilde{f}_{h,s} \cdot \nabla$. We note that if f is analytic in a complex neighborhood of x and the one-step integrator (10) is a Runge-Kutta method, then the coefficients $d_j(x)$ of (12) and the functions $\tilde{f}_j(x)$ of the modified Eq. (13) are also analytic [18, Chap. 9]. The expressions (16) and (17) then read $\phi(\varphi_t(x)) = e^{t\mathcal{L}_D}\phi(x)$ and $\phi(\tilde{\varphi}_t(x)) = e^{t\widetilde{\mathcal{L}}_D}\phi(x)$, respectively. If a numerical method has order p, then it holds

$$e^{h\mathcal{L}_D}\phi(x) - \phi(\psi_h(x;f)) = \mathcal{O}(h^{p+1})$$

and by definition of the modified Eq. (13) we have

[2]Formally in backward error analysis one considers an infinite series $\tilde{f}_h(x) = f(x) + h\tilde{f}_1(x) + h^2\tilde{f}_2(x) + \cdots$ such that $\widetilde{X}(t_n) = X_n$ at $t_n = nh$. But the infinite series for $\tilde{f}_h$ is usually not convergent and one needs therefore to consider an appropriate truncation.

$$e^{h\widetilde{\mathcal{L}}_D}\phi(x) - \phi(\psi_h(x;f)) = \mathcal{O}(h^{s+2}).$$

2.2 Integrator Based on Modified Equations

The approach of modifying integrators that first appeared in [11] consists in using a modified differential equation in order to construct higher order numerical integrators while preserving geometric properties like symplecticity. Consider an integrator (10) of order p for the problem (9). The idea to derive a more precise numerical integrator [still for the problem (9)] is to construct a modified differential equation for $s \geq p$

$$\begin{aligned} \frac{d\hat{X}}{dt} &= \hat{f}_{h,s}(\hat{X}) \\ \hat{X}(0) &= X_0, \end{aligned} \tag{18}$$

where

$$\hat{f}_{h,s}(x) = f(x) + h^p \hat{f}_p(x) + \cdots + h^s \hat{f}_s(x) \tag{19}$$

such that when applying the integrator (10)–(18) it has order $s+1$ for the original Eq. (9) that is,

$$\psi_h(x;\hat{f}_{h,s}) - \varphi_h(x) = \mathcal{O}(h^{s+2}),$$

or in terms of Lie derivatives

$$\phi(\psi_h(x;\hat{f}_{h,s})) - e^{h\mathcal{L}_D}\phi(x) = \mathcal{O}(h^{s+2}).$$

The functions $\hat{f}_i(x)$ can be uniquely defined by the condition that $X(h) = \psi_h(x;\hat{f}_{h,s}) + \mathcal{O}(h^{s+2})$ for $h \to 0$ and it then follows that $X(t_n) = \psi_h(X_{n-1};\hat{f}_{h,s}) + \mathcal{O}(h^{s+1})$ for $h \to 0$ and bounded times $t_n = nh \leq T$.

3 A Framework for Constructing and Analyzing Stochastic Integrators

We recall that associated to the SDE (1), there exists a differential operator $\mathcal{L}$, called the generator of the SDE, defined by

$$\mathcal{L} := f \cdot \nabla + \frac{1}{2} g g^T : \nabla^2, \tag{20}$$

where $\nabla^2 \phi$ denotes the Hessian of a function ϕ [scalar product on matrices are denoted $A : B = \text{trace}(A^T B)$]. Then the function $u(t,x) = \mathbb{E}\,(\phi(X(t))|X_0 = x)$ is the solution of the partial differential equation, called the backward Kolmogorov equation, given by

$$\frac{\partial u}{\partial t} = \mathcal{L} u, \qquad u(x,0) = \phi(x), \tag{21}$$

where $\phi \in \mathcal{V}^\infty(E, \mathbb{R})$. Using a Taylor expansion together with Eq. (21) gives a series for $u(t,x)$ in terms of the generator of the SDE [12, 41]

$$u(x,h) - \phi(x) = \sum_{j=1}^{l} \frac{h^j}{j!} \mathcal{L}^j \phi(x) + h^{l+1} r_l(f,g,\phi)(x), \tag{22}$$

where under appropriate smoothness on f, g, ϕ the remainder $r_l(f,g,\phi)$ has polynomial growths ($E = \mathbb{R}^d$) or can be bounded ($E = \mathbb{T}^d$). We next consider for the numerical method (2) the function

$$U(x,h) = \mathbb{E}(\phi(X_1)|X_0 = x), \tag{23}$$

and assume that it can be expanded as

$$U(x,h) = \phi(x) + h A_0(f,g)\phi(x) + h^2 A_1(f,g)\phi(x) + \cdots, \tag{24}$$

where $A_i(f,g),\ i = 0,1,2,\ldots$ are linear differential operators that depend on the choice of the integrator with coefficients depending on f, g, and their derivatives. Assume also that for all $i = 0,1,2,\ldots$

$$A_i(f + \eta\hat{f}, g + \eta\hat{g}) = A_i(f,g) + \eta A_i(f,\hat{f},g,\hat{g}) + \mathcal{O}(\eta^2), \tag{25}$$

where $A_i(f,\hat{f},g,\hat{g}) + \mathcal{O}(\eta^2)$ is again a differential operator. As [see (5)] we have $|\mathbb{E}(\phi(X_1)) - \mathbb{E}(\phi(X(t_1)))| = |U(x,h) - u(x,h)|$ for all methods of weak order $p \geq 1$ we must have $A_0 = \mathcal{L}$ (consistency condition). Furthermore, a method of weak local order $p \geq 1$ satisfies

$$|\mathbb{E}(\phi(X_1)) - \mathbb{E}(\phi(X(t_1)))| = h^{p+1} \left(A_p - \frac{\mathcal{L}^{p+1}}{(p+1)!} \right) \phi(x) + \mathcal{O}(h^{p+2}). \tag{26}$$

A global weak order result can also be expressed in terms of the above differential operators [30, Chap. 2.2, 2.3]. Indeed we have

Theorem 1 *Assume that f, g in (1) are C^∞ with bounded derivatives up to any order and consider a numerical integrator (2) on $[0,T]$ with an expansion of the form (24) with bounded moments $\mathbb{E}|X_n|^\kappa$, $\kappa \in \mathbb{N}$ sufficiently large. Assume that the numerical integrator has weak local order p with a constant $C = C(x)$ with*

polynomial growth. Then, we have the following expansion of the global error, for all $\phi \in \mathcal{V}^\infty(E, \mathbb{R})$,

$$\mathbb{E}(\phi(X(T))) - \mathbb{E}(\phi(X_N)) = h^p \int_0^T \mathbb{E}(\psi_e(X(s), s))ds + \mathcal{O}(h^{p+1}), \tag{27}$$

where $Nh = T$ *and* $\psi_e(x, t)$ *satisfies*

$$\psi_e(x, t) = \left(A_p - \frac{1}{(p+1)!}\mathcal{L}^{p+1}\right) u(x, t), \tag{28}$$

and $u(x, t)$ *is the solution of* (21).

3.1 High Weak Order Method Based on Modified Equations

We observe that an expansion of the type (24) holds for many numerical integrators and hence (24) is not a restrictive assumption.

Example 1 We start with an example and consider the SDE (1) with $d = m = 1$ and define for $X_0 = x$ the (semi-implicit) θ-Milstein method by

$$\begin{aligned} X_{n+1} &= X_n + (1-\theta)hf(X_n) + \theta hf(X_{n+1}) + g(X_n)\Delta W_n \\ &\quad + \frac{1}{2}g'(X_n)g(X_n)((\Delta W_n)^2 - h), \end{aligned} \tag{29}$$

where the Wiener increment ΔW_n are given by independent $\mathcal{N}(0, h)$ random variables. Taylor expansion of (23) reveals that $U(x, h) = \phi(x) + hA_0(f, g)\phi(x) + h^2A_1(f, g)\phi(x) + \mathcal{O}(h^3)$ with

$$A_1(f, g)\phi(x) = \theta\left[f'(x)f(x) + \frac{1}{2}f''(x)g^2(x)\right]\phi'(x)$$

$$+\frac{1}{2}\left[f^2(x) + 2\theta f'(x)g^2(x) + \frac{1}{2}(g'(x)g(x))^2\right]\phi''(x) \tag{30}$$

$$+\frac{1}{2}\left[g'(x)g^3(x) + g^2f(x)\right]\phi'''(x) + \frac{1}{8}g^4(x)\phi^{(4)}(x). \tag{31}$$

An easy computation then shows that

$$\left(\frac{1}{2}\mathcal{L}^2\phi - A_1(f, g)\phi\right)(x) = \left(\frac{1}{2} - \theta\right)\left(f'(x)f(x) + \frac{1}{2}f''(x)g^2(x)\right)\phi'(x)$$

$$+\left(\left(\frac{1}{2} - \theta\right)f'(x)g(x) + \frac{1}{2}g'(x)f(x) + \frac{1}{4}g^2(x)g''(x)\right)g(x)\phi''(x).$$

In view of Theorem 1 we can deduce that this integrator has weak order one and applying the integrator to a simple particular SDE (e.g. linear) reveals that it will not be second order in general. Note that for $\theta = 0$ we recover the Euler-Maruyama (EM) method $X_{n+1} = X_n + hf(X_n) + g(X_n)\Delta W_n$ that has also weak order one. However, the more involved θ-Milstein method has better strong and mean-square stability behavior for $\theta > 0$ [19].

Construction of higher order weak methods based on a basic classical integrator for example (29) can be achieved by using the framework of modified equations summarized in Sect. 2. Consider the following modified SDE based on (1)

$$d\hat{X} = f_{h,s}(\hat{X})dt + g_{h,s}(\hat{X})dW(t), \quad \hat{X}(0) = X_0, \tag{32}$$

where

$$f_{h,s}(x) = f(x) + hf_1(x) + \cdots + h^s f_s(x), \tag{33}$$

$$g_{h,s}(x) = g(x) + hg_1(x) + \cdots + h^s g_s(x). \tag{34}$$

Inserting the expansion for $f_{h,s}, g_{h,s}$ into the generator of the SDE (32) given by

$$\hat{\mathcal{L}}\phi := f_{h,s} \cdot \nabla_x \phi + \frac{1}{2}(g_{h,s} g_{h,s}^T) : \nabla_x^2 \phi, \tag{35}$$

yields

$$\hat{\mathcal{L}} = \mathcal{L} + h\mathcal{L}_1 + h^2\mathcal{L}_2 + \cdots + h^s\mathcal{L}_s + \mathcal{O}(h^{s+1}), \tag{36}$$

where for $j = 1, 2, \ldots s$ the operators $\mathcal{L}_j$ is given by

$$\mathcal{L}_j = f_j \cdot \nabla_x + \frac{1}{2}\sum_{k=0}^{j}(g_k g_{j-k}^T) : \nabla_x^2, \tag{37}$$

using the notations $f_0 := f$ and $g_0 := g$. We will sometimes write $\mathcal{L}_j(f_j, g, g_1, \ldots, g_j)$ to emphasize on which functions the coefficients of the operator depend. Going back to the Example 1, we set $p = s = 1$ in the modified Eq. (33) and we obtain using Assumption (25) for the θ-Milstein method applied to (32) the expansion

$$\begin{aligned} U(x,h) &= \phi(x) + h\hat{\mathcal{L}}\phi(x) + h^2 A_1(f_{h,1}, g_{h,1})\phi(x) + \cdots, \\ &= \phi(x) + h\mathcal{L} + h^2(\mathcal{L}_1 + A_1(f,g))\phi(x) + \mathcal{O}(h^3). \end{aligned}$$

Hence to obtain a second order weak integrator for the SDE (1) we must in view of Theorem 1 define f_1, g_1 in the modified equations such that

$$\mathcal{L}_1 = \left(\frac{1}{2}\mathcal{L}^2 - A_1\right),$$

and in view of (30) we easily find that

$$f_1(x) = \left(\frac{1}{2} - \theta\right) f'(x)f(x) + \frac{1}{2}\left(\frac{1}{2} - \theta\right) f''(x)g^2(x),$$
$$g_1(x) = \left(\frac{1}{2} - \theta\right) f'(x)g(x) + \frac{1}{2}g'(x)f(x) + \frac{1}{4}g^2(x)g''(x).$$

hence the integrator

$$X_{n+1} = X_n + (1-\theta)hf_{h,1}(X_n) + \theta hf_{h,1}(X_{n+1}) + g_{h,1}(X_n)\Delta W_n + \frac{1}{2}g'(X_n)g(X_n)((\Delta W_n)^2 - h). \tag{38}$$

has weak order two when applied to a one-dimensional SDE (1). A priori the last term of the above integrator should read $\frac{1}{2}g'_{h,1}(X_n)g_{h,1}(X_n)((\Delta W_n)^2 - h)$ but it can be shown (direct calculation) that

$$\mathbb{E}\left(g'_{h,1}(X_n)g_{h,1}(X_n)((\Delta W_n)^2 - h)\right) = \mathbb{E}(g'(X_n)g(X_n)((\Delta W_n)^2 - h)) + \mathcal{O}(h^3), \tag{39}$$

and hence we can substitute $g_{h,1}$ with g in the last term of the modified integrator without altering the weak order two of the method.

The scheme (39) can easily be generalized to multidimensional SDEs (see [2]). We note that in general, constructing second order weak methods require to solve many order conditions (system of equations) if relying on Itô-Taylor expansion [36]. The integrator (38) derived in [2] belongs to a class of general weak second order integrators derived originally by Milstein [29]. For $\theta = 0$ Talay proved second order convergence in [37] and for $\theta = 1/2$, the scheme was shown to have favorable stability properties for scalar SDEs with additive noise [19]. For $\theta = 1$, the method seems to have first appeared in [2] where it is also shown that it possesses good mean-square stability properties for scalar SDEs with multiplicative noise.

A general result for high order weak integrator based on modified equations can be obtained recursively as follows: consider a numerical integrator (2) of order p for the SDE (1) and assume that a modified Eq. (33) with $s = p + r - 2$ has been obtained so that (2) applied to (32) is an integrator of weak order $p + r - 1$ for the original SDE (1), i.e.,

$$\begin{aligned} U(x,h) &= \phi(x) + hA_0(f_{h,p+r-2}, g_{h,p+r-2}) + \cdots \\ &\quad + h^{p+r-1}A_{p+r-2}(f_{h,p+r-2}, g_{h,p+r-2}) + \mathcal{O}(h^{p+r+1}) \\ &= \phi(x) + h\mathcal{L} + \cdots + \frac{h^{p+r-1}}{(p+r-1)!}\mathcal{L}^{p+r-1} + h^{p+r}R(f,g). \end{aligned}$$

Consider now the SDE (1) and a modified equation with $s = p + r - 1$

$$f_{h,p+r-1}(x) = f_{h,p+r-2}(x) + h^{p+r-1} f_{p+r-1}(x),$$
$$g_{h,p+r-1}(x) = g_{h,p+r-2}(x) + h^{p+r-1} g_{p+r-1}(x),$$

then in view of the above expansion for $U(x,h)$ we obtain

$$\begin{aligned} U(x,h) &= \phi(x) + hA_0(f_{h,p+r-1}, g_{h,p+r-1}) + \cdots + h^{p+r} A_{p+r-1}(f_{h,p+r-1}, g_{h,p+r-1}) \\ &\quad + \mathcal{O}(h^{p+r+1}) \\ &= \phi(x) + h\mathcal{L} + \cdots + \frac{h^{p+r-1}}{(p+r-1)!}\mathcal{L}^{p+r-1} \\ &\quad + h^{p+r}(\mathcal{L}_{p+r-1}(f_{p+r-1}, g, g_1, \ldots, g_{p+r-1}) + R(f,g)) + \mathcal{O}(h^{p+r+1}), \end{aligned}$$

where we used the assumption (25), the equality $A_0(f_{h,p+r-1}, g_{h,p+r-1}) = \hat{\mathcal{L}}(f_{h,p+r-1}, g_{h,p+r-1})$, the expansion (36) and the fact that integrator is of order $p+r-1$ for the original SDE (1). We see that if we can find f_{p+r-1}, g_{p+r-1} such that the differential operator $\mathcal{L}_{p+r-1}$ satisfies

$$\mathcal{L}_{p+r-1}(f_{p+r-1}, g, g_1, \ldots, g_{p+r-1}) = \frac{1}{(p+r)!}\mathcal{L}^{p+r} - R(f,g)$$

then according to Theorem 1, the integrator (2) will be of order $p+r$ for the original SDE (1).

The framework presented above and introduced in [2] has been further used to construct new high weak order methods that are mean square stable and new high weak order invariant preserving stochastic methods. In Sects. 3.2, 4.1, and 4.2 we will further explain how this framework can be used to construct new high order methods for the approximation of invariant measure of ergodic SDEs following [4, 5].

3.2 *High Order Numerical Approximation of the Invariant Measure of Ergodic SDEs*

We now explain how ideas from backward error analysis and modified equations can lead to efficient approximations of the invariant measure of ergodic SDEs. We assume that the SDE (1) is ergodic (see Assumption 3.1 or 3.2) and that its unique invariant measure μ has a density function ρ_∞. We recall that ρ_∞ is then the unique solution of the Fokker-Planck equation

$$\mathcal{L}^* \rho_\infty = 0, \tag{40}$$

where $\mathcal{L}^*\phi = -\nabla\cdot(\phi f) + \frac{1}{2}gg^T : \nabla^2\phi$ is the L^2-adjoint of the generator $\mathcal{L}$ defined in (20). We consider a numerical integrator (2) that we assume to be ergodic and to have an expansion of the form (24) satisfying (25). To motivate our approach, consider a numerical integrator of weak order p. Passing to the limit $T \to \infty$ in (27) we obtain for all $\phi \in \mathcal{V}^\infty(E,\mathbb{R})$ and $h \to 0$,

$$e(\phi, h) = \lambda_p h^p + \mathcal{O}(h^{p+1}), \tag{41}$$

where $e(\phi,h)$ is defined in (8), for any deterministic initial condition, with λ_p defined as

$$\lambda_p = \int_0^{+\infty}\int_{\mathbb{R}^d}\left(A_p - \frac{1}{(p+1)!}\mathcal{L}^{p+1}\right)u(y,t)\rho_\infty(y)dydt, \tag{42}$$

where $u(x,t)$ is the solution of (21). Hence an ergodic numerical integrator of weak order p is also of order p with respect to the invariant measure of (1). Next using the L^2-adjoint of the differential operator $\frac{1}{(p+1)!}\mathcal{L}^{p+1} - A_p$ reveals

$$\lambda_p = -\int_0^{+\infty}\int_{\mathbb{R}^d} u(y,t)\left(A_p^* - \frac{1}{(p+1)!}(\mathcal{L}^*)^{p+1}\right)\rho_\infty(y)dydt.$$

Now from (40), $(\mathcal{L}^*)^{p+1}\rho_\infty = 0$, hence if in addition to have weak order p the numerical integrator also satisfies $A_p^*\rho_\infty = 0$ then it will have order $p+1$ with respect to the invariant measure. Of course if it happens that the integrator is of weak order $p+1$, then necessarily $A_p^*\rho_\infty = 0$, since in that case $\mathcal{L}^{p+1}/(p+1)! = A_p$. It turns out that that there exist integrators of weak order p but not $p+1$ that still fulfill $A_p^*\rho_\infty = 0$.

We now discuss a generalization of the above result and derive order conditions for constructing numerical integrators that approximate the invariant measure of ergodic SDEs with high order. The starting point is again the expansion (24) for the numerical method. We also derive a modified generator

$$\widetilde{\mathcal{L}} = \mathcal{L} + \sum_{i\geq 1} h^i L_i, \tag{43}$$

but we now require, following the framework of backward error analysis, that its expansion match (formally) the expansion of the numerical method, i.e.,

$$U(x,h) - \phi(x) = \sum_{j\geq 1}\frac{h^j}{j!}\widetilde{\mathcal{L}}^j\phi(x).$$

Using the above equality and matching equal power of h one obtains using (24) [12, 41]

$$L_n = A_n - \frac{1}{2}(\mathcal{L}L_{n-1} + L_{n-1}\mathcal{L} + \cdots) - \cdots - \frac{1}{(n+1)!}\mathcal{L}^{n+1}. \tag{44}$$

Compared to the framework for ODEs recalled in Sect. (2) we face an additional difficulty when trying to apply the ideas of backward error analysis in the SDE context. Indeed, in view of (13), one would be tempted to truncate the formal expansion for $\widetilde{\mathcal{L}}$, say $\widetilde{\mathcal{L}}_N = \mathcal{L} + \sum_{i=1}^{N} h^i L_i$, and to consider the backward Kolmogorov equation

$$\frac{\partial \tilde{u}_N}{\partial t} = \widetilde{\mathcal{L}}_N \tilde{u}_N.$$

However, the existence of a solution to the above PDE is not clear as $\widetilde{\mathcal{L}}_N$ is no longer a second order operator in general. The proper definition of the Eq. (43) is based on the following result. Under appropriate assumptions on f, g, assume further there exists a constant λ and for all integer $k \geq 0$ constants C_k, κ_k such that for all $t \geq 0$

$$\|u(t,\cdot) - \int_{E^d} \phi(y)\rho_\infty(y)dy\|_{C^k} \leq C_k(1 + t^{\kappa_k})e^{-\lambda t}\|\phi\|_{C^k}, \tag{45}$$

where $\|v(t,\cdot)\|_{C^k}$ denotes the sup norm of the function $v(x,t)$ and its derivatives with respect to x up to order k. Then it has been shown in [12] ($E = \mathbb{T}$) and in [22] ($E = \mathbb{R}$) that for all $\ell \in \mathbb{N}$, there exist smooth functions $\tilde{u}_\ell(t,x)$ defined for all $t \geq 0$ that satisfies for all $N \in \mathbb{N}$

$$\frac{\partial \tilde{u}_N}{\partial t} - L\tilde{u}_N = \sum_{i=1}^{N} L_i v_{N-1}.$$

This result is used in the following lemma that is central in our numerical framework. We will consider two distinct situations either

Assumption 3.1 *$E = \mathbb{T}^d$ and*

- *f, g are C^∞ functions on the torus $\mathbb{T}^d$;*
- *the generator $\mathcal{L}$ is elliptic or hypo-elliptic;*
- *in the case where $\mathcal{L}$ is hypo-elliptic, we further assume the uniqueness of the invariant measure of (1).*

or

Assumption 3.2 *$E = \mathbb{R}^d$ and*

- *f, g are of class C^∞, with bounded derivatives of any order, and g is bounded;*
- *the generator $\mathcal{L}$ in (20) is a uniformly elliptic operator,* i.e. *there exists $\alpha > 0$ such that for all $x, \xi \in \mathbb{R}^d$, $x^T g(\xi) g(\xi)^T x \geq \alpha x^T x$;*
- *there exist $C, \beta > 0$ such that for all $x \in \mathbb{R}^d$, $x^T f(x) \leq -\beta x^T x + C$.*

We note that under either of the above assumptions, the SDE (1) has a unique invariant measure. For the case $E = \mathbb{R}^d$, (3.2) also implies that the density function of ρ_∞ of the invariant measure has bounded moments of any order, i.e., for all $n \geq 0$

$$\int_{\mathbb{R}^d} |x|^n \rho_\infty(x)dx < \infty. \tag{46}$$

The following lemma is valid either for the torus $E = \mathbb{T}^d$ [12] under Assumption 3.1 or for $E = \mathbb{R}^d$ under Assumption 3.2 [22].

Lemma 2 *Consider L_n the operators defined in (44). Then there exists a sequence of functions $(\rho_n(x))_{n\geq 0}$ such that $\rho_0 = \rho_\infty$ and for all $n \geq 1$, $\int_E \rho_n(x)dx = 0$ and*

$$\mathcal{L}^* \rho_n = -\sum_{l=1}^{n} (L_l)^* \rho_{n-l}. \tag{47}$$

For any positive integer N, setting

$$\rho_N^h(x) = \rho_\infty(x) + \sum_{n=1}^{N} h^n \rho_n(x), \tag{48}$$

then there exists a constant $C(N, \phi)$ such that for all $\phi \in \mathcal{V}^\infty(E, \mathbb{R})$

$$\left| \int_{E^d} \phi(x) d\mu^h(x) - \int_{E^d} \phi(x) \rho_N^h(x) dx \right| \leq C(N, \phi) h^{N+1}, \tag{49}$$

where $C(N, \phi)$ is independent of h.

We have seen in the beginning of this section if in addition to have weak order p the numerical integrator also satisfies $A_p^* \rho_\infty = 0$ then it can achieve weak order $p+1$ with respect to the invariant measure. We now assume that a numerical integrator has an expansion (24) with differential operators A_j satisfying

$$A_j^* \rho_\infty = 0, \quad \text{for} \quad j = 1, \ldots r-1. \tag{50}$$

If (50) holds, then using Lemma 2 we can show that the numerical method has order r with respect to the invariant measure. For example, assume $A_1^* \rho_\infty = 0$ then by Lemma 2 $\mathcal{L}^* \rho_1 = -L_1^* \rho_\infty = (A_1^* - \frac{1}{2}(\mathcal{L}^*)^2)\rho_\infty = 0$ and using (48) with $N = 1$ we obtain $\rho_1^h(x) = \rho_\infty(x) + \mathcal{O}(h^2)$ and using (49) we see that we obtain a numerical method of order 2 for the invariant measure. By induction, we have the following theorem [4].

Theorem 2 *Suppose that the SDE (1) satisfies Assumptions 3.1 or 3.2. Consider an ergodic numerical method satisfying assumptions (24) and (25). Assume that (50) holds. Then the numerical integrator has (at least) order r in (8) for the invariant measure, i.e., for all $\phi \in \mathcal{V}^\infty(E, \mathbb{R})$*

$$e(\phi, h) = h^r \int_0^\infty \int_{\mathbb{T}^d} A_r u(x, t) \rho_\infty(x) dx dt + \mathcal{O}(h^{r+1})$$

$$= -h^r \int_0^\infty \int_{\mathbb{T}^d} u(x,t) A_r^* \rho_\infty(x) dx dt + \mathcal{O}(h^{r+1}),$$

where $u(x,t)$ solves the backward Kolmogorov equation (21).

3.2.1 Construction of High Order Numerical Approximation of Invariant Measure

Following Theorem 2, the task is now to construct a numerical method (24) such that (50) holds. Of course a sufficient condition to fulfill (50) is by choosing a method of weak order r. But this is not necessary as we will show here for a class of SDEs with invariant measures of the form

$$\rho_\infty(x) = Z e^{-V(x)} \tag{51}$$

where $Z = (\int_{E^d} e^{-V(x)} dx)^{-1}$ is a normalization constant, and V is a smooth function of class C^∞. The construction is based on modified equations.

Theorem 3 *Consider an ergodic system of SDEs (1) with an invariant measure of the form (51) satisfying Assumption 3.1 or (3.2). Consider a numerical method satisfying Assumptions (24) and (25) of order p for the invariant measure. Then, for all fixed $m \geq 1$, there exists a modified SDE of the form*

$$dX = \left(f + h^p f_p + \cdots + h^{p+m-1} f_{p+m-1}\right) dt + g dW \tag{52}$$

such that the numerical method applied to this modified SDE satisfies

$$A_j^* \left(f + h^p f_p + \cdots + h^{p+m-1} f_{p+m-1}, g\right) \rho_\infty = 0 \quad j = p, \ldots, p+m-1. \tag{53}$$

Furthermore, if the numerical method applied to the modified SDE is ergodic, then it yields a method of order (at least) $r = p + m$ in (8) for the invariant measure of (1).

We observe in the above theorem that a modified Eq. (32) involving only the drift term (33) is used. This theorem can be proved by induction. We sketch the ideas of the construction of the modified equation and refer to [5] for details. Assume that $f_j, j < k < p + m$ have been constructed and consider the scheme obtained by applying the numerical method to the modified SDE

$$dX = \left(f + \cdots + h^{k-1} f_{k-1}\right) dt + g dW$$

so that the numerical integrator (24) applied to this modified SDE satisfies (53) for $j < k$. Using integration by part and the form of the differential operator A_j it can be shown that

$$\int_{E^d} (A_k\phi)\rho_\infty dx = \int_{E^d} (\widetilde{A}_k\phi)\rho_\infty dx, \quad \text{for all } \phi \in \mathcal{V}^\infty(E,\mathbb{R}), \tag{54}$$

where $A_k = (f + \cdots + h^{k-1}f_{k-1}, g)$, $\widetilde{A}_k$ is of the form $\widetilde{A}_k = -F \cdot \nabla$, for a certain F. From (54), we deduce

$$A_k^* \left(f + \cdots + h^{k-1}f_{k-1}, g\right) \rho_\infty = \mathrm{div}(F\rho_\infty). \tag{55}$$

We set $f_k := F$. Since $A_0 = \mathcal{L}$ we have

$$A_0^* \left(f + \cdots + h^{k-1}f_{k-1} + h^k f_k, g\right) \phi = A_0^* \left(f + \cdots + h^{k-1}f_{k-1}, g\right) \phi - h^k \mathrm{div}(f_k\phi),$$

using (24), (25), and (55) we obtain

$$\begin{aligned} &A_k^* \left(f + \cdots + h^{k-1}f_{k-1} + h^k f_k, g\right) \rho_\infty \\ &\quad = A_k^* \left(f + \cdots + h^{k-1}f_{k-1}, g\right) \rho_\infty - \mathrm{div}(f_k\rho_\infty) = 0. \end{aligned}$$

Together with an induction argument, this shows (53) and using Theorem 2, we conclude that the scheme applied to the modified SDE (52) has order $p + m$ for the invariant measure.

4 Construction of High Order Numerical Methods for Ergodic Dynamical Systems

In this section we discuss the actual construction of high order numerical methods for ergodic dynamical systems. We will focus on special yet important classes of SDEs, namely Brownian dynamics that describe the motion of a particle in a potential subject to thermal noise and Langevin dynamics that models the motion of a particle in a potential subject to linear friction and molecular diffusion [34]. In both cases, we will use $E = \mathbb{R}^d$. We mention the recent work [40] that introduces post-processing techniques for SDEs combined with ideas of modified equations [4] that also allows to construct higher order methods for ergodic SDEs (see [9] for an extension to SPDEs).

4.1 Brownian Dynamics

The Brownian dynamics is described by the following SDE

$$dX(t) = -\nabla V(X(t))dt + \sigma dW(t), \tag{56}$$

where $V : \mathbb{R}^d \to \mathbb{R}$ is a smooth potential, $\sigma > 0$ is a constant, and $W = (W_1, \ldots, W_d)^T$ is a standard d-dimensional Wiener process. The invariant measure of this SDE, assuming ergodicity, is given by the Gibbs density function

$$\rho_\infty(x) = Ze^{-2V(x)/\sigma^2}, \tag{57}$$

where Z is a normalization constant. To illustrate the above theory, we first consider the special case $d = m = 1$ and will mention the generalization later. As a basic integrator, we consider the θ-method

$$X_{n+1} = X_n + (1-\theta)hf(X_n) + \theta hf(X_{n+1}) + \sqrt{h}\sigma\xi_n, \tag{58}$$

where $\xi_n \sim \mathcal{N}(0,1)$ are independent dimensional Gaussian random variables. This methods as weak order 1 when $\theta \neq 1/2$ and weak order two for $\theta = 1/2$. In particular for $\theta = 0$ it collapse has to the well-known Euler-Maruyama (EM) method. Note that for the SDE (56), $f = -V'(x)$. We now construct a method or order two for the invariant measure that is only of weak order one in general. For the method (58), a direct calculation reveals that $A_0 = \mathcal{L}$ (due to the weak order one) and

$$A_1\phi = \frac{1}{2}f^2\phi'' + \frac{\sigma^2}{2}f\phi''' + \frac{\sigma^4}{8}\phi^{(4)} + \theta\left(f'f\phi' + \frac{\sigma^2}{2}f''\phi' + \sigma^2 f'\phi''\right).$$

Integration by part in (54) shows that

$$\widetilde{A}_1\phi = \left(-(1-2\theta)\left(\frac{1}{2}f'f + \frac{\sigma^2}{4}f''\right)\right)\phi'$$

hence according to Theorem 3 defining $f_1 = -(1-2\theta)\left(\frac{1}{2}f'f + \frac{\sigma^2}{4}f''\right)$ and applying (58) to the modified equation $dX = (f + hf_1)dt + \sigma dW$ will produce a second order method for the approximation of the invariant measure of (56). We note that for linear one-dimensional problems (56) the method (58) with $\theta = 1/2$ samples exactly the invariant measure (see [26]), hence it is not surprising that $f_1 = 0$ in that case.

For the multi-dimensional case (56), we can go through the same derivation by setting this time $f = -\nabla V(x)$. One obtains $f_1 = -(1-2\theta)\left(\frac{1}{2}f'f + \frac{\sigma^2}{4}\Delta f\right)$ and the scheme

$$X_{n+1} = X_n + (1-\theta)h(f + hf_1)(X_n) + \theta h(f + hf_1)(X_{n+1}) + \sqrt{h}\sigma\xi_n, \tag{59}$$

with $\xi_{n,i} \simeq \mathcal{N}(0,1), i = 1, 2, \ldots, d$ will be of second order for the invariant measure of (56).

4.1.1 Removing Derivatives in Integrators Based on Modified Equations: Runge-Kutta Formulation

In general numerical integrator based on modified equations need to evaluate derivatives of drift or diffusion functions. Sometimes such derivatives are cheap to compute [11]. When such a computation is not convenient, the derivatives appearing in these integrator can also be aproximated by "finite differences" introducing internal stages. For example, consider the modified Euler-Maruyama scheme obtained from the above modified θ method with $\theta = 0$ (multi-dimensional case)

$$X_{n+1} = X_n + h\left(f(X_n) - h\frac{1}{2}f'f(X_n) + \frac{\sigma^2}{4}\Delta f(X_n)\right) + \sqrt{h}\sigma\xi_n. \tag{60}$$

One can check that the derivative free version

$$\begin{aligned} Y_1 &= X_n + \sqrt{2}\sigma\sqrt{h}\xi_n \\ Y_2 &= X_n - \frac{3}{8}hf(Y_1) + \frac{\sqrt{2}}{4}\sigma\sqrt{h}\xi_n \\ X_{n+1} &= X_n - \frac{1}{3}hf(Y_1) + \frac{4}{3}hf(Y_2) + \sigma\sqrt{h}\xi_n, \end{aligned} \tag{61}$$

$\xi_{n,i} \simeq \mathcal{N}(0,1), i = 1, 2, \ldots, d$ is also of weak order 1, while approximating with second order the invariant measure of (56). This can be seen by checking that the same operators A_0, A_1 appear in the expansion (24) of both schemes.

We close this section by a numerical experiment with the above derivative free second order method for the invariant measure. We consider a Brownian dynamics (56) with a two-dimensional quartic potential $V(x) = (1 - x_1^2)^2 + (1 - x_2^2)^2 + \frac{x_1 x_2}{2} + \frac{x_2}{5}$. We emphasize that doing so we depart from the case of Lipschitz vector fields for which our theory apply. We will see numerically that we still get the right order of the invariant measure under these weaker assumptions. The Gibbs invariant density function is depicted in Fig. 1 (left picture). We consider the Euler-Maruyama method [(58) with $\theta = 0$] and the second order modified Euler Maruyama method (61). We see in Fig. 1 (right picture) that the modified method captures the invariant measure with the rate predicted by Theorem 3.

4.2 Langevin Dynamics

We consider here the Langevin equation given by a second order stochastic differential equation of the form

$$dq(t) = M^{-1}p(t)dt, \tag{62a}$$

$$dp(t) = (-\nabla V(q(t)) - \gamma p(t))\, dt + \sqrt{2\beta^{-1}\gamma} M^{1/2} dW(t), \tag{62b}$$

where $p(t), q(t) \in \mathbb{R}^d$, $W(t)$ denotes a standard d-dimensional Wiener process, $V : \mathbb{R}^d \to \mathbb{R}$ is a potential, M is a symmetric positive definite mass matrix (taken as the identity in what follows for simplicity), and the positive scalar parameter γ, β are related, respectively, to friction and temperature. The dynamics generated by (62) is ergodic provided suitable smoothness and growth assumptions on the Hamiltonian energy (see, e.g., [21, 28]),

$$H(p, q) = \frac{1}{2} p^T p + V(q), \tag{63}$$

with invariant measure given by the Gibbs density function [see e.g., (57)]

$$\rho_\infty(p, q) = Z e^{-\beta H(p,q)}, \tag{64}$$

where Z is the normalization constant. Efficient numerical integrators for Langevin equations are based on the Lie-Trotter splitting [5, 7, 23, 25]

$$X_{n+1} = \Phi_h \circ \Theta_{h,n}(X_n), \tag{65}$$

where $X_n = (p_n, q_n)^T$. The integrator Φ_h approximates the exact flow of the deterministic Hamiltonian part

$$dq(t) = p(t)dt, \qquad dp(t) = -\nabla V(q(t))dt, \tag{66}$$

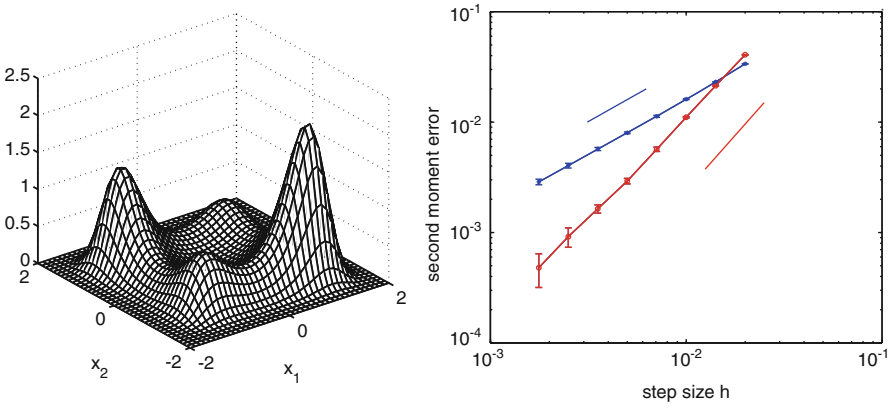


Fig. 1 Figures adapted from [4]. *Left picture*: Brownian dynamics with quartic potential. Gibbs density function. *Right picture*: Error $e(\phi, h)$ defined in (8) for $\phi(x) = x_1^2 + x_2^2$ for the EM method (*blue line with star symbol*) and the modified EM method (*red line with circle symbol*). Standard deviation is depicted by the *vertical line* obtained from the Monte Carlo error originating from ten trajectories. The *straight lines without symbols* represent first and second order slopes

while $\Theta_{h,n}$ is an integrator for the stochastic part given by

$$q_{n+1} = q_n, \qquad p_{n+1} = e^{-\gamma h} p_n + \sqrt{\beta^{-1}(1 - e^{-2\gamma h})}\xi_n,$$

where $\xi_n \sim \mathcal{N}(0, I)$ are independent d-dimensional Gaussian random variables. This integrator has the same law of probability as the exact solution of the stochastic part that is given by the variation of constants formula,

$$q(t_{n+1}) = q(t_n), \qquad p(t_{n+1}) = e^{-\gamma h} p(t_n) + \sqrt{2\beta^{-1}\gamma} \int_{t_n}^{t_{n+1}} e^{-\gamma(h-s)} dW(s).$$

For simplicity we assume that the potential V is a $\mathcal{C}^\infty$ function where ∇V has bounded derivatives of any order and satisfies standard growth condition $q^T \nabla V(q) \geq C_1 q^T q - C_2$, for all $q \in \mathbb{R}^d$ for $C_1, C_2 > 0$, which guarantee that (62) is ergodic and we assume that the numerical flow Φ_h is globally Lipschitz in $\mathbb{R}^{2d}$. In the non-globally Lipschitz case, one could resort to implicit deterministic integrator to avoid exploding trajectories as studied in [21] or apply the theory presented below for explicit integrators while rejecting exploding trajectories. This latter procedure that has been rigorously analyzed in [31] and applied to ergodic Langevin problems in [32].

Consider the numerical integrator $X_{n+1} = \Phi_h \circ \Theta_{h,n}(X_n)$ with an expansion (24). According to Theorem 2, if (50) holds, then the numerical integrator will approximate the invariant measure of (62) with order r. To construct accurate integrators with respect to the invariant measure we will as in Sect. 3.2 consider a modified equation. As here the only integrator to be considered is a deterministic integrator Φ_h, we consider the modified deterministic ODE (13) $s = r$, where $f(t) = (p(t), -\nabla V(q(t)))^T$. The question now is how the condition $A_j^* \rho_\infty = 0$ of Theorem 2 translates into a condition on the $\hat{f}_j$ in (19) ?

The connection can be revealed by using the semi-group property of the Markov process

$$\mathbb{E}(\phi(X_1)|X_0 = x) = \mathbb{E}(\phi(\Phi_h \circ \Theta_{h,n})(X_0)|X_0 = x) = e^{h\mathcal{L}_S}(\phi \circ \Phi_h)(x), \qquad (67)$$

for a smooth test function ϕ and $x \in \mathbb{R}^{2d}$, where $e^{h\mathcal{L}_S}\phi$ denotes the exact flow of the Kolmogorov backward equation corresponding to the stochastic part of (62) with generator $\mathcal{L}_S$ given by

$$\mathcal{L}_S := -\gamma p \cdot \nabla_p + \beta^{-1}\gamma \Delta_p.$$

We then have in view of (16) with $M = r$

$$\mathbb{E}(\phi(X_1)|X_0 = x) = \left(\sum_{k=0}^{r} \frac{h^k \mathcal{L}_S^k}{k!}\right)\left(\sum_{k=0}^{r} \frac{h^k \widetilde{\mathcal{L}}_D^k}{k!}\right)\phi(x) + \mathcal{O}(h^{r+1})$$

$$= \phi(x) + h\mathcal{L}\phi(x) + \sum_{k=1}^{r} h^{k+1} A_k \phi(x) + \mathcal{O}(h^{r+1}),$$

where we recall that $\widetilde{\mathcal{L}}_D = \tilde{f}_{h,r} \cdot \nabla := F_0 + hF_1 + \cdots + F_r$, where $F_j = f_j \cdot \nabla$, $j = 1, 2, \ldots r$ and $f_0 = f$. Developing the sums in the above equality and identifying power of h allows to find an expression for A_k in terms of power of $\mathcal{L}_S$ and product of operators F_j. It is then deduced that if

$$\operatorname{div}(f_j \rho_\infty) = 0, \qquad j = 1, \ldots, r-1, \tag{68}$$

then $A_j^* \rho_\infty = 0$, $j = 1, \ldots, r-1$ and $A_r^* \rho_\infty = \operatorname{div}(f_r \rho_\infty)$. Using Theorem 2 we find

Theorem 4 *Assume that Φ_h is a consistent method for (66) with Lipschitz continuous flow. If the vector fields in (13) ($s = r$) satisfy (68) with $r \geq 1$, then assuming ergodicity, the Lie-Trotter splitting (65) has order r of accuracy for the invariant measure of (62), i.e.,*

$$e(\phi, h) = -h^r \int_0^\infty \int_{\mathbb{R}^d \times \mathbb{R}^d} u(p, q, t) \operatorname{div}(f_r(p, q) \rho_\infty(p, q)) dp dq dt + \mathcal{O}(h^{r+1}), \tag{69}$$

for all $\phi \in \mathcal{C}_P^\infty(\mathbb{R}^d, \mathbb{R})$ and $h \to 0$, where $e(\phi, h)$ is defined in (8) and $u(x, t)$ is the solution of the Backward Kolmogorov equation (21) for (62).

The condition (68) can be rewritten

$$\operatorname{div}(f_j \rho_\infty) = \big(\operatorname{div}(f_j) - \beta f_j \cdot \nabla H\big) \rho_\infty.$$

Observe that $\operatorname{div}(f_j) = 0$ for all $1 \leq j \leq r-1$ is equivalent to the fact that the deterministic integrator Φ_h is volume preservation up to order r, i.e., $\det(\partial \Phi_h(y)/\partial y) = 1 + \mathcal{O}(h^{r+1})$. Also $f_j \cdot \nabla H = 0$ for all $1 \leq j \leq r-1$ is equivalent to the fact that the deterministic integrator Φ_h is energy preserving up to order r, i.e., $H(\Phi_h(y)) = H(y) + \mathcal{O}(h^{r+1})$. Notice that any deterministic method of order r will fulfill both conditions and hence will produce a Lie-Trotter splitting method of order r for the invariant measure of (62) according to Theorem 4. In [7] it was shown that sufficient conditions to preserve the invariant measure of (62) up to order r is to consider a symplectic integrator in the Lie-Trotter splitting preserving the energy with order r. The condition (68), first given in [5], is thus a weaker characterization of high order Lie-Trotter splitting for Langevin dynamics. We also mention the work [23, 25] where efficient non-Markovian schemes with second order accuracy for the invariant measure of (62) have been constructed.

As an illustration of the above theory we consider three different deterministic numerical integrators, namely the explicit Euler method,

$$p_{n+1} = p_n - h\nabla V(q_n), \qquad q_{n+1} = q_n + hp_n,$$

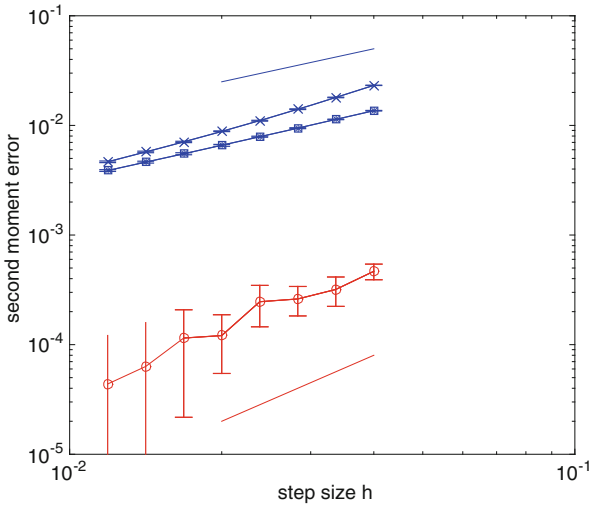


Fig. 2 Figures adapted from [5]. Langevin dynamics with quartic potential : error $e(\phi, h)$ defined in (8) for $\phi(p, q) = p^2 + q^2$ and with deterministic flow in the Lie-Trotter splitting given by the EM method (*blue line with cross symbol*), the symplectic Euler method (*blue line with square symbol*), and the Heun method (*red line with circle symbol*). Standard deviation is depicted by the *vertical line* obtained from the Monte Carlo error originating from ten trajectories. The *straight lines without symbols* represent first and second order slopes

the symplectic Euler method,

$$p_{n+1} = p_n - h\nabla V(q_n), \qquad q_{n+1} = q_n + hp_{n+1},$$

and the Heun method, a second order explicit method given by

$$p_{n+1} = p_n - h\nabla V\left(q_n + \frac{h}{2}p_n\right), \qquad q_{n+1} = q_n + h\left(p_n - \frac{h}{2}\nabla V(q_n)\right).$$

Departing from the situation of globally Lipschitz vector fields (for which the theory has been derived) we consider the Langevin dynamics (62) with a quartic potential $V(q) = (1 - q^2)^2 - \frac{1}{2}q$. As for the previous case of Brownian dynamics we observe numerically that the theory still apply and Fig. 2 corroborate the results of Theorem 4.

5 Conclusion

We have discussed the use of techniques originating from geometric integration such as backward error analysis and modified equations for the construction of stochastic integrators. We have focused in particular on algorithms for the computation of

expectation of functionals and invariant measures of stochastic processes. This approach for the construction and analysis of stochastic integrators sheds new light on the interplay between weak approximation of stochastic differential equations and accurate computation of their invariant measures. In particular, new sufficient conditions to approximate the invariant measure of an ergodic system independently of the underlying weak order of the method have been described. A better understanding of the numerical approximation of ergodic SDEs is also important for applications in biology, chemistry, or physics, where efficient computation of expectation of functionals of a stochastic process is used for sampling or free energy calculations.

References

1. A. Abdulle, G. Pavliotis, Numerical methods for stochastic partial differential equations with multiple scales. J. Comput. Phys. **231**(6), 2482–2497 (2012)
2. A. Abdulle, D. Cohen, G. Vilmart, K.C. Zygalakis, High order weak methods for stochastic differential equations based on modified equations. SIAM J. Sci. Comput. **34**(3), 1800–1823 (2012)
3. A. Abdulle, W. E, B. Engquist, E. Vanden-Eijnden, The heterogeneous multiscale method. Acta Numer. **21**, 1–87 (2012)
4. A. Abdulle, G. Vilmart, K.C. Zygalakis, High order numerical approximation of the invariant measure of ergodic SDEs. SIAM J. Numer. Anal. **52**(4), 1600–1622 (2014)
5. A. Abdulle, G. Vilmart, K.C. Zygalakis, Long time accuracy of Lie-Trotter splitting methods for Langevin dynamics. SIAM J. Numer. Anal. **53**(1), 1–16 (2015)
6. A. Abdulle, G. Pavliotis, U. Vaes, Spectral methods for multiscale stochastic differential equations. SIAM/ASA J. Uncertain. Quantif. (2016, to appear)
7. N. Bou-Rabee, H. Owhadi, Long-run accuracy of variational integrators in the stochastic context. SIAM J. Numer. Anal. **48**(1), 278–297 (2010)
8. C.-E. Bréhier, Analysis of an HMM time-discretization scheme for a system of stochastic PDEs. SIAM J. Numer. Anal. **51**(2), 1185–1210 (2013)
9. C.-E. Bréhier, G. Vilmart, High-order integrator for sampling the invariant distribution of a class of parabolic stochastic PDEs with additive space-time noise. SIAM J. Sci. Comput. **38**, A2283–A2306 (2016)
10. Y. Cao, D.T. Gillespie, L. Petzold, The slow scale stochastic simulation algorithm. J. Chem. Phys. **122**, 014116 (2005)
11. P. Chartier, E. Hairer, G. Vilmart, Numerical integrators based on modified differential equations. Math. Comp. **76**(260), 1941–1953 (2007) (electronic)
12. A. Debussche, E. Faou, Weak backward error analysis for SDEs. SIAM J. Numer. Anal. **50**(3), 1735–1752 (2012)
13. W. E, D. Liu, E. Vanden-Eijnden, Analysis of multiscale methods for stochastic differential equations. Commun. Pure Appl. Math. **58**(11), 1544–1585 (2005)
14. W. E, D. Liu, E. Vanden-Eijnden, Nested stochastic simulation algorithms for chemical kinetic systems with multiple time scales. J. Comput. Phys. **221**(1), 158–180 (2007)
15. C.W. Gardiner, *Handbook of Stochastic Methods. For Physics, Chemistry and the Natural Sciences*, 2nd edn. (Springer, Berlin, 1985)
16. D. Gillespie, Stochastic simulation of chemical kinetics. Annu. Rev. Phys. Chem. **58**, 35–55 (2007)
17. C. Graham, D. Talay, Stochastic simulation and Monte Carlo methods, in *Stochastic Modelling and Applied Probability*. Mathematical Foundations of Stochastic Simulation, vol. 68 (Springer, Heidelberg, 2013)

18. E. Hairer, C. Lubich, G. Wanner, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Series in Computational Mathematics, vol. 31, 2nd edn. (Springer, Berlin, 2006)
19. D.J. Higham, A-stability and stochastic stability mean-square stability. BIT Numer. Math. **40**, 404–409 (2000)
20. P. Kloeden, E. Platen, *Numerical Solution of Stochastic Differential Equations* (Springer, Berlin, 1992)
21. M. Kopec, Weak backward error analysis for Langevin process. BIT Numer. Math. **55**(4), 1057–1103 (2015)
22. M. Kopec, Weak backward error analysis for overdamped Langevin processes. IMA J. Numer. Anal. **35**(2), 583–614 (2015)
23. B. Leimkuhler, C. Matthews, Rational construction of stochastic numerical methods for molecular sampling. Appl. Math. Res. Express **2013**, 34–56 (2013)
24. B. Leimkuhler, C. Matthews, Molecular dynamics, in *Interdisciplinary Applied Mathematics,* vol. 36 (Springer, Cham, 2015), pp. xxii+443
25. B. Leimkuhler, C. Matthews, G. Stoltz, The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. IMA J. Numer. Anal. **36**(1), 13–79 (2016)
26. T. Li, A. Abdulle, W. E, Effectiveness of implicit methods for stiff stochastic differential equations. Commun. Comput. Phys. **3**(2), 295–307 (2008)
27. J. Mattingly, A. Stuart, D. Higham, Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. Stochastic Process. Appl. **101**(2), 185–232 (2002)
28. J.C. Mattingly, A.M. Stuart, M.V. Tretyakov, Convergence of numerical time-averaging and stationary measures via poisson equations. SIAM J. Numer. Anal. **48**(2), 552–577 (2010)
29. G. Milstein, Weak approximation of solutions of systems of stochastic differential equations. Theory Probab. Appl. **30**(4), 750–766 (1986)
30. G. Milstein, M. Tretyakov, *Stochastic Numerics for Mathematical Physics*. Scientific Computing (Springer, Berlin, 2004)
31. G.N. Milstein, M.V. Tretyakov, Numerical integration of stochastic differential equations with nonglobally Lipschitz coefficients. SIAM J. Numer. Anal. **43**(3), 1139–1154 (2005) (electronic)
32. G.N. Milstein, M.V. Tretyakov, Computing ergodic limits for Langevin equations. Phys. D **229**(1), 81–95 (2007)
33. G. Pavliotis, A. Stuart, *Multiscale Methods: Averaging and Homogenization*. Text in Applied Mathematics, vol. 53 (Springer, New York, 2008)
34. H. Risken, *The Fokker-Planck Equation*. Springer Series in Synergetics, vol. 18 (Springer, Berlin, 1989)
35. G.O. Roberts, R.L. Tweedie, Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli **2**(4), 341–363 (1996)
36. A. Rößler, Second order Runge-Kutta methods for Itô stochastic differential equations. SIAM J. Numer. Anal. **47**(3), 1713–1738 (2009)
37. D. Talay, *Efficient Numerical Schemes for the Approximation of Expectations of Functionals of the Solution of a SDE and Applications*. Lecture Notes in Control and Information Sciences, vol. 61 (Springer, Berlin, 1984), pp. 294–313
38. D. Talay, Second order discretization schemes of stochastic differential systems for the computation of the invariant law. Stochast. Stochast. Rep. **29**(1), 13–36 (1990)
39. E. Vanden-Eijnden, Numerical techniques for multiscale dynamical system with stochastic effects. Commun. Math. Sci. **1**, 385–391 (2003)
40. G. Vilmart, Postprocessed integrators for the high order integration of ergodic SDEs. SIAM J. Sci. Comput. **37**(1), A201–A220 (2015)
41. K.C. Zygalakis, On the existence and the applications of modified equations for stochastic differential equations. SIAM J. Sci. Comput. **33**(1), 102–130 (2011)

Stability and Strong Convergence for Spatial Stochastic Kinetics

Stefan Engblom

Mathematics Subject Classification (2010). 60J27, 60J28, 92C42.

1 Introduction

Since complete microscopic descriptions of biochemical processes inside a living cell is a nearly hopeless task, stochastic *mesoscopic* models have emerged as important tools in Systems Biology and Neuroscience. Formulated in a way which resembles the familiar macroscopic ODE/PDE-based viewpoint, but with randomness accounting for microscopic effects such as thermal movements and molecular noise, *mesoscopic* stochastic models attempt to strike a balance between computational feasibility and accuracy. Similar models have been used also in Epidemiology and Population studies, where instead randomness is included to make up for the missing details in irregular or unknown contact patterns and in the varying environment of individual agents.

Assuming that the possible presence of stochastic effects is motivating the choice of modeling framework, analyzing model stability is a crucial issue that is often overlooked. In particular, when designing numerical or approximate models of some kind, it is extremely important that the resulting model somehow inherits the stability of the original mathematical model. Without such bridging results, the use of in silico support for phenomena observed under stochastic nonlinear effects and operating under model uncertainty can rightly be questioned.

S. Engblom (✉)
Division of Scientific Computing, Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden
e-mail: stefane@it.uu.se

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_5

Quite often the discussion of stability issues and well-posedness tend to take either a more mathematical standpoint or be of a more, say, intuitive character. I like to argue here that a fair dose of mathematical rigor is needed and that it can actually help in building an intuitive grasp of the modeling process itself. However, I would also like to express the perhaps provocative thought that the most general situation in a mathematical framework, while complete and concentrated in content, can also be difficult to connect with actual applications.

Monographs containing general mathematical treatments of SDEs include [33, Chap. V], [20, Chap. 11], [5, Chap. 6], and [35, Chaps. 3–5], where various types of jump-diffusion SDEs are discussed in a non-spatial context. A study of the stochastic flow of scalar jump SDEs in continuous state is found in [32] and the existence of solutions to SPDEs driven by Poisson processes is treated in [25].

In applied contexts, well-posedness is often thought to automatically follow from the physical situation at hand. Consequently solutions to stochastic models are simply assumed to remain inside some bounded region of state-space [36, Chap. V]. Mathematically, the problem is that for non-trivial stochastic models operating under *open* conditions, *any* state will be reached with a non-zero probability. Oftentimes a preliminary analysis of such open models under the assumption of a priori bounded solutions can be cast into a rigorous stopping time argument. But as we shall see, usually the associated error estimates are lost and only convergence properties remain. Examples of this kind of numerical analysis for various time discretization methods are found in [2–4, 15, 26, 27, 30].

In the current review we will summarize some of the ideas for well-posedness and stability of stochastic reaction networks found in [17], later put to use in the context of numerical analysis [12, 18]. Similar ideas also emerged around the same time in [11, 34], and all of these are in fact related to the earlier theory in [31]. A contribution with the present study is to re-cast those results in a spatial context.

So far the main body of numerical and multiscale analysis has been done in the traditional well-stirred setting (with some notable exceptions [6, 37]). In [19] a general method for bringing the mesoscopic stochastic description into a spatially extended model on arbitrary geometries was described, resulting also in a highly general software [14]. The computational complexity of spatial models is considerably much higher compared to the well-stirred case, and effective computational methods are therefore desirable. The requirement for a consistent numerical analysis in this setting is the starting point for the present discussion.

This review is structured as follows. In Sect. 2 we investigate more closely the purpose with understanding and analyzing stability in modeling in general, and in particular within the present context of numerical analysis of stochastic models. In Sect. 3 we summarize the mathematical machinery required for spatially extended continuous-time Markov chains. The stability analysis is presented in Sect. 4 and consists of, firstly, a brief discussion of some technical tools together with a suitable set of modeling assumptions, and secondly, an existence and uniqueness result together with a basic exemplifying perturbation result. A concluding discussion is found in Sect. 5.

2 Meaning and Use of Stability

The concept of stability in computational modeling has a somewhat context-dependent meaning, but typically involves the development of a priori regularity bounds for the solution of the mathematical model considered and some kind of continuity estimate over the input data. For stochastic models, bounds in the absolute mean, or mean square sense have become popular choices in the numerical analysis community. For a time-dependent stochastic process $X(t)$ in some state space this means bounds of the kind

$$\mathbb{E}[\|X(t)\|^p] \text{ or } \mathbb{E}\left[\sup_{s\in[0,t]}\|X(t)\|^p\right] \leq B(t, X_0), \quad p \in \{1, 2\}, \tag{1}$$

where $B(t, X_0)$ depends on time and on the regularity of the initial data X_0. With $Y(t)$ a solution of a perturbed version of the model, a similar bound on the difference process $(Y - X)(t)$ is also desirable. For if the model is perturbed slightly, we expect that also the solution changes in relation to the magnitude of the perturbation. Note that this concept requires the randomness of the two processes X and Y to be suitably coupled according to some prescription. Depending on the context, this coupling can in fact be rather involved and require some effort [1, 7]. Finally, we mention also a somewhat different kind of stability property which we will not discuss further here, namely *ergodicity*, which handles the limit $t \to \infty$, e.g. in (1).

When deciding to use a stochastic model, there are usually some reasons beforehand to believe that randomness has a potentially important impact on the system studied. In fact, the very purpose of the modeling could well be to study the effects of the presence of the *aleatory uncertainty*, that is, the process' inherent physical randomness.

This implies that, on the computational side, we are faced with a mathematical model thought to operate under conditions where stochasticity has non-trivial effects, and the computational analysis is therefore often concerned with the behavior around particularly intriguing parameter combinations. This clearly motivates a careful afterthought when designing computational methods. When the original model itself is sensitive to perturbations of various kinds, hastily constructed approximations may easily lead to incorrect results.

2.1 The Lax Principle

The celebrated *Lax principle*, first formulated in the context of difference approximations for PDEs in the mid-50s [29], states that consistent schemes for well-posed problems converge if and only if they are stable. The astonishing generality of this principle can be appreciated as follows [16].

Consider metric spaces (X, d_X) and (Y, d_Y), and an operator $T : X \longrightarrow Y$. The mathematical problem $Tx = y$ for some given $y \in Y$ is *well-posed* if T^{-1} is continuous in some neighborhood of the data y. Similarly, a sequence of numerical approximations $T_n x_n = y$ is *stable* if $(T_n^{-1})_{n\geq 1}$ all exist and are continuous in a neighborhood of y. *Consistency*, finally, says that $T_n x \longrightarrow Tx$ for x in a sufficiently large subset of X. *Such an approximation is convergent*.

For any $\delta > 0$, there is an $N \geq 1$ such that by consistency, $d_Y(T_n x, Tx) = d_Y(T_n x, T_n x_n) < \delta$ whenever $n > N$. Put $y_n = T_n x$. Given an $\varepsilon > 0$ we can find a δ such that by stability, $d_X(x, x_n) = d_X(T_n^{-1} y_n, T_n^{-1} y) < \varepsilon$ whenever $d_Y(y_n, y) < \delta$, as implied by $n > N$. This is the same thing as convergence $x_n \longrightarrow x$ as $n \longrightarrow \infty$.

The overall recipe can be summarized as follows: first show that both the original and the approximating model share similar stability characteristics. Then show that the approximation is uniformly consistent in residual space. Finally conclude convergence and, if possible, develop a bound for the error.

We now proceed to show how to implement these abstract steps in the concrete setting when T represents a spatially extended continuous-time Markov chain and when y is the data of the model.

3 Discrete-in-Space Markovian Kinetics

3.1 Master Equations

Through the pioneering work of Gillespie [24] stochastic simulation techniques became a popular tool when studying the kinetics of reaction networks at the discrete single molecule level. The actual model is a continuous-time Markov chain (CTMC), commonly described via the *chemical master equation* (CME) [23]. Let the state $x \in \mathbf{Z}_+^D = \{0, 1, 2, \ldots\}^D$ count the number of molecules of each of D species and let R reactions be prescribed as transitions of the state, $x \to x - \mathbb{N}_r$, where each $\mathbb{N}_r \in \mathbf{Z}^D$ is a transition step. Then the CME can be written as

$$\frac{\partial p(x,t)}{\partial t} = \sum_{r=1}^{R} w_r(x + \mathbb{N}_r)p(x + \mathbb{N}_r, t) - w_r(x)p(x,t) =: \mathcal{M}p(x,t),$$

that is, an equation of state for the D-dimensional probability density of the system conditioned upon some initial state. Here each *reaction propensity*, $w_r : \mathbf{Z}_+^D \to \mathbf{R}_+$, governs the probability per unit of time for reaction r to occur.

The CME is derived under the condition of a homogeneous molecular distribution in the domain considered. This assumption is violated when the transport of molecules through the solvent is slow or more generally, when concentration gradients may build up [13, 21]. As a viable modeling approach, the diffusion at a molecular level can be treated as a special set of reactions yielding the reaction-diffusion master equation (RDME) [23]. Specifically, let the state-vector

$x \in \mathbf{Z}_+^{D\times N}$, where N is the number of compartments required to discretize the considered volume. Then for two neighboring compartments i and j, the transition $(x_{d,i}, x_{d,j}) \to (x_{d,i} - 1, x_{d,j} + 1)$ corresponds to a transport event of species d from compartment i to j. The RDME can formally be written as

$$\frac{\partial p(x,t)}{\partial t} = (\mathcal{M} + \mathcal{D})p(x,t),$$

where $\mathcal{M}$ is the reaction-, and $\mathcal{D}$ the transport (e.g. diffusion) operator, respectively. The dimensionality of the state-space is now much higher and sampling even single trajectories is a computationally intensive problem.

3.2 Pathwise Representations

Consider first a space-independent state variable $X(t) \in \mathbf{Z}_+^D$ counting at time t the number of entities among D species. The state transitions $X \mapsto X - \mathbb{N}_r$ are prescribed probabilistically by

$$\mathbf{P}\left[X(t+dt) = x - \mathbb{N}_r \middle| X(t) = x\right] = w_r(x)\,dt + o(dt), \tag{2}$$

for $r = 1, \ldots, R$. Assuming a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ supporting R-dimensional Poisson processes, the state is evolved according to [20]

$$X_i(t) = X_i(0) - \sum_{r=1}^{R} \mathbb{N}_{ri} \Pi_r \left(\int_0^t w_r(X(s))\,ds \right), \tag{3}$$

for species $i = 1, \ldots, D$ and with standard unit-rate and independent Poisson processes Π_r. For brevity, in the present contribution we will exclusively remain in the operational time framework (3). A corresponding jump SDE representation may also be developed, see [1, 15, 17, 30] for examples.

When (3) takes place on the nodes of a spatial network, a notation for the connecting transport process needs to enter. A given volume V_{tot} may be discretized into J smaller voxels such that the state $X \in \mathbf{Z}_+^{D\times J}$, where in this context X_{ij} is the number of molecules of the ith species in the jth voxel. The dynamics (3) is used anew, but on a per-voxel basis. Adding a general transport process we get

$$\begin{aligned} X_{ij}(t) = X_{ij}(0) &- \sum_{r=1}^{R} \mathbb{N}_{ri} \Pi_{rj} \left(\int_0^t w_{rj}(X_{\cdot j}(s))\,ds \right) \\ &- \sum_{k=1}^{J} \Pi'_{ijk} \left(\int_0^t q_{ijk} X_{ij}(s)\,ds \right) + \sum_{k=1}^{J} \Pi'_{ikj} \left(\int_0^t q_{ikj} X_{ik}(s)\,ds \right), \end{aligned} \tag{4}$$

where q_{ijk} is the rate per unit of time for species i to move from the jth voxel to the kth. This type of linear transport rate covers many traditional physical processes, like diffusion and convection [19], but can also be augmented to handle sub-diffusion [9] and general models of active transport [23]. Importantly, the network implied by (4) need not depend on a discretized form of a transport operator, but can also be obtained by incorporating observed transport patterns [8].

Equation (4) consists of two parts: a model for the *local physics* expressed with a set of general nonlinear laws, and a connecting *global transport process*, which is conservative due to the two opposing terms. The model relates to stochastic partial differential equations (SPDEs) in that they both govern a stochastic and spatially extended process. However, an SPDE is formulated in continuous space and time, with time typically discretized first in a numerical scheme. In (4) the model is rather directly formulated in an already discretized space. What makes (4) attractive as a framework for computational modeling is the great expressivity as well as the fact that effective numerical algorithms may be devised.

4 Stability and Convergence

4.1 Tools and Assumptions

An important technique in stochastic analysis is the *stopping time*. Define $\tau_P := \inf_{t\geq 0}\{\|X(t)\| > P\}$ in some suitable norm and for $P > 0$. The *stopped process* is $X(\hat{t}) = X(t\wedge\tau_P)$ and the idea is that analytic results developed for $X(\hat{t})$ may be lifted to the "untamed" original process $X(t)$. The technique most often used is based on *Fatou's lemma*, which states that

$$\mathbb{E}[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} \mathbb{E}[X_n]. \tag{5}$$

Specifically, suppose one can show the stopped bound $\mathbb{E}[\|X(\hat{t})\|] \leq B(t)$ and that the goal is to transfer this bound to the original process.

Firstly, for any $\omega \in \Omega$, $X(s,\omega)$ either explodes for some $0 \leq s \leq t$, or is bounded by some value P_0. In the former case we have that $\|X(\hat{s},\omega)\|$ is unbounded as $P \to \infty$. In the latter case, $\lim_{P\to\infty} \|X(\hat{s},\omega)\| = \|X(s,\omega)\|$ for $0 \leq s \leq t$.

Since $\mathbb{E}[\|X(\hat{t})\|]$ is bounded from above independently of P (this is the assumed bound) we must have that $\lim_{P\to\infty} \|X(\hat{t})\| = \|X(t)\|$ almost surely; in other words, $\|X(t)\|$ cannot diverge to ∞ at a non-zero probability.

Now for a convergent sequence, the limit and limit inferior are equal, and hence $\lim_{P\to\infty} \|X(\hat{t})\| = \liminf_{P\to\infty} \|X(\hat{t})\|$ almost surely, and so we have

$$\mathbb{E}[\lim_{P\to\infty} \|X(\hat{t})\|] = \mathbb{E}[\liminf_{P\to\infty} \|X(\hat{t})\|], \tag{6}$$

where Fatou's lemma finally applies.

The integral representation (4) and the stopping time imply another important tool in analysis, namely *Dynkin's formula* [10, Chap. 9.2.2]. For $f : \mathbf{Z}_+^{D\times J} \to \mathbf{R}$ a suitable function,

$$\begin{aligned}
&\mathbb{E}\left[f(X(\hat{t})) - f(X(0))\right] = \\
&\quad \mathbb{E}\left[\int_0^{\hat{t}} \sum_{j=1}^{J}\sum_{r=1}^{R} w_{rj}(X_{\cdot j}(s))\left[f(X(s) - \mathbb{N}_r \mathbb{1}_j^T) - f(X(s))\right] ds\right] \\
&\quad + \mathbb{E}\left[\int_0^{\hat{t}} \sum_{j,k=1}^{J}\sum_{i=1}^{D} q_{ijk} X_{ij}(s)\left[f(X(s) - \mathbb{1}_i\mathbb{1}_j^T + \mathbb{1}_i\mathbb{1}_k^T) - f(X(s))\right] ds\right] \\
&\quad + \mathbb{E}\left[\int_0^{\hat{t}} \sum_{j,k=1}^{J}\sum_{i=1}^{D} q_{ikj} X_{ik}(s)\left[f(X(s) + \mathbb{1}_i\mathbb{1}_j^T - \mathbb{1}_i\mathbb{1}_k^T) - f(X(s))\right] ds\right],
\end{aligned} \tag{7}$$

where $\mathbb{1}_j$ is an all-zero vector with a single 1 at position j. Below we frequently generalize a bound derived from Dynkin's formula in stopped time by using Fatou's lemma.

In spatial modeling the transition intensities in (4) are generally *density dependent*, meaning that $w_{rj}(x) = V_j u_r(V_j^{-1}x)$ for some dimensionless function u_r [20, Chap. 11]. Further, since we are dealing with counting processes, stability bounds are naturally expressed in a weighted norm. We thus assume the existence of a suitable vector $\boldsymbol{l}$ with $\min_i \boldsymbol{l}_i = 1$ and define

$$\|x\|_{\boldsymbol{l}} := \boldsymbol{l}^T x, \quad x \in \mathbf{R}_+^D. \tag{8}$$

Following [12, 17] closely we formulate our assumptions as follows:

Assumption 4.1 (Running Assumptions) For a mesh M consisting of voxel volumes $(V_j)_{j=1}^J$ we assume that

$$w_{rj}(x) = V_j u_r(V_j^{-1}x), \text{ where } u \text{ is independent of the mesh,} \tag{9}$$

$$-\boldsymbol{l}^T \mathbb{N} u(x) \le A + \alpha \|x\|_{\boldsymbol{l}}, \tag{10}$$

$$(-\boldsymbol{l}^T \mathbb{N})^2 u(x)/2 \le B + \beta_1 \|x\|_{\boldsymbol{l}} + \beta_2 \|x\|_{\boldsymbol{l}}^2, \tag{11}$$

$$|u_r(x) - u_r(y)| \le L_r(P)\|x - y\|, \text{ for } r = 1 \ldots R, \text{ and } \|x\|_{\boldsymbol{l}} \vee \|y\|_{\boldsymbol{l}} \le P, \tag{12}$$

$$m_V \bar{V}_M \le V_j \le M_V \bar{V}_M, \text{ for average voxel volume } \bar{V}_M, \tag{13}$$

$$\forall_{ij}\ |\{k;\ q_{ijk} \ne 0\}| \le M_D. \tag{14}$$

Equations (10) and (11) put bounds on the process in the direction measured by $\boldsymbol{l}$, while (13) and (14) assure that the mesh is not too far from uniform and with a bounded connectivity.

We now proceed to combine Dynkin's formula (7) and Assumption 4.1 to develop stability results for the process governed by (4).

4.2 Existence and Uniqueness

For spatially varying solutions we consider the following generalization of (8),

$$\|x\|_{\boldsymbol{l},1} \equiv \sum_{j=1}^{J} \|x_{\cdot j}\|_{\boldsymbol{l}} = \boldsymbol{l}^T x \mathbf{1}, \quad x \in \mathbf{R}_+^{D\times J}, \tag{15}$$

with $\mathbf{1}$ an all-unit vector. From Dynkin's formula (7) we find

$$\mathbb{E}\left[\|X(\hat{t})\|_{\boldsymbol{l},1}^p\right] = \mathbb{E}\left[\|X(0)\|_{\boldsymbol{l},1}^p\right] + \mathbb{E}\left[\int_0^{\hat{t}} F(X(s))\,ds\right] \tag{16}$$

where

$$F(X) \equiv \sum_{j=1}^{J}\sum_{r=1}^{R} w_{rj}(X_{\cdot j})\left[\left(\|X\|_{\boldsymbol{l},1} - \boldsymbol{l}^T\mathbb{N}_r\right)^p - \|X\|_{\boldsymbol{l},1}^p\right]. \tag{17}$$

This shows a feature with the norm (15); the transport process is conservative and therefore does not affect the norm.

The following two inequalities will come in handy [17, Lemma 4.6]: let $f(x) \equiv (x+y)^p - x^p$ with $x \in \mathbf{R}_+$ and $y \in \mathbf{R}$. Then for integer $p \geq 1$,

$$f(x) \leq pyx^{p-1} + C_p y^2\left[x^{p-2} + |y|^{p-2}\right], \tag{18}$$

$$|f(x)| \leq C_p|y|\left[x^{p-1} + |y|^{p-1}\right], \tag{19}$$

where $C_p > 0$ only depends on p and may be different on each occasion of use.

Using (18), assumptions (13) and (9)–(11) we obtain

$$F(X) \leq p(AV_{\text{tot}} + \alpha x)x^{p-1} + C_p(BV_{\text{tot}} + \beta_1 x + \beta_2^V x^2)(x^{p-2} + C_{\mathbb{N}}^{p-2}), \tag{20}$$

where for brevity $x \equiv \|X\|_{l,1}$ and where $\beta_2^V := \beta_2 m_V^{-1} \bar{V}_M^{-1}$, and $C_{\mathbb{N}} := \|\boldsymbol{l}^T \mathbb{N}\|_\infty$. Combining (16) and (20) we readily obtain a bound of the form

$$\mathbb{E}\left[\left\|X(\hat{t})\right\|_{l,1}^p\right] \leq \mathbb{E}\left[\|X(0)\|_{l,1}^p\right] + \mathbb{E}\left[\int_0^t C(1 + \|X(\hat{s})\|_{l,1}^p)\, ds\right], \tag{21}$$

for some $C > 0$. By Grönwall's inequality we arrive at a P-independent bound for $\mathbb{E}\left[\left\|X(\hat{t})\right\|_{l,1}^p\right]$. Letting $P \to \infty$ and invoking Fatou's lemma we deduce the following

Theorem 4.2 (Moment Bound, [12, Theorem 2.2]) *Let $X(t)$ obey (4) under Assumption 4.1. Then for any integer $p \geq 1$,*

$$\mathbb{E}\left[\|X(t)\|_{l,1}^p\right] \leq \left(\mathbb{E}\left[\|X(0)\|_{l,1}^p\right] + 1\right) \exp(Ct) - 1, \tag{22}$$

where the constant $C > 0$ depends on p and on the constants in the assumptions.

Note that when the mesh contains small voxels V_j and when $\beta_2 > 0$ in (11), then C for $p \geq 2$ in (21) contains the term $\beta_2 V_j^{-1}$, indicating a possibly rapid growth of moments.

While the moment bound in Theorem 4.2 implies a certain degree of regularity, an even stronger result is required to achieve path-wise control in the form of strong continuity with respect to perturbations or to numerical parameters. This can be achieved via Burkholder's inequality (see (26) below) provided the *quadratic variation* of the process is controlled. This is defined by

$$[Y]_t = \operatorname*{plim}_{\|\mathcal{P}\| \to 0} \sum_{k=0}^{n-1} \left(Y_{t_{k+1}} - Y_{t_k}\right)^2, \tag{23}$$

for $Y(t)_{t \geq 0}$ a real-valued process and a time partition $\mathcal{P} = \{0 = t_0 < t_1 < \cdots < t_n = t\}$ where $\|\mathcal{P}\| := \max_k |t_{k+1} - t_k|$.

Lemma 4.3 (Quadratic Variation, [12, Lemma 2.3]) *Let $X(t)$ satisfy (4) under Assumption 4.1. Then the quadratic variation of $\|X(t)\|_{l,1}^p$ is bounded by*

$$\mathbb{E}\left[[\|X\|_{l,1}^p]_t^{1/2}\right] \leq \mathbb{E}\left[\int_0^t C(1 + \|X(s)\|_{l,1}^p + \beta_2^V \|X(s)\|_{l,1}^{p+1})\, ds\right], \tag{24}$$

where $C > 0$ depends on p and on the constants in Assumption 4.1, but not on the mesh resolution, but where $\beta_2^V := \beta_2 m_V^{-1} \bar{V}_M^{-1}$.

Proof By writing $Y(t) = \|X(t)\|^p$ in (23) as a Lebesgue-Stieltjes integral and using the inequality $\|\cdot\|_2 \le \|\cdot\|_1$ one arrives at the bound (see [12, Lemma 2.3] for details)

$$\mathbb{E}\left[[\|X\|_{l,1}^p]_{\hat{t}}^{1/2}\right] \le \mathbb{E}\left[\int_0^{\hat{t}} \sum_{j=1}^{J}\sum_{r=1}^{R} w_{rj}(X_{\cdot j}(s)) \left|\left(\|X(s)\|_{l,1} - \boldsymbol{l}^T\mathbb{N}_r\right)^p - \|X(s)\|_{l,1}^p\right| ds\right].$$

Using this time the inequality (19) and assumptions (9) and (11),

$$\le \mathbb{E}\left[\int_0^{\hat{t}} \sum_{j,r} C_p|\boldsymbol{l}^T\mathbb{N}_r|w_{rj}(X_{\cdot j}(s))\left[\|X(s)\|_{l,1}^{p-1} + |\boldsymbol{l}^T\mathbb{N}_r|^{p-1}\right] ds\right]$$

$$\le \mathbb{E}\left[\int_0^{\hat{t}} C_p(BV_{\text{tot}} + \beta_1\|X(s)\|_{l,1} + \beta_2^V\|X(s)\|_{l,1}^2)(\|X(s)\|_{l,1}^{p-1} + C_{\mathbb{N}}^{p-1})\, ds\right].$$

After expanding and using some simple bounds we may, in view of Theorem 4.2, let $P \to \infty$ to arrive at (24). □

As a straightforward generalization of the construction used in [17] we shall consider the following class of path-wise locally bounded processes:

$$S_{\mathcal{F}}^{p,\text{loc}}(\mathbf{Z}_+^{D\times J}) = \left\{X(t,\omega) : \begin{array}{l} X(t) \in \mathbf{Z}_+^{D\times J} \text{ is } \mathcal{F}_t\text{-adapted such that} \\ \mathbb{E}[\sup_{t\in[0,T]}\|X_t\|_{l,1}^p] < \infty \text{ for } \forall T < \infty \end{array}\right\}. \tag{25}$$

Theorem 4.4 (Existence, [12, Theorem 2.4]) *Let $X(t)$ be a solution to (4) under Assumptions 4.1 with $\beta_2 = 0$. Then if $\mathbb{E}[\|X(0)\|_{l,1}^p] < \infty$, $\{X(t)\}_{t\ge0} \in S_{\mathcal{F}}^{p,\text{loc}}(\mathbf{Z}_+^{D\times J})$. If $\beta_2 > 0$ then the conclusion remains under the additional requirement that $\mathbb{E}[\|X(0)\|_{l,1}^{p+1}] < \infty$.*

As mentioned the proof relies on *Burkholder's* inequality [33, Chap. IV.4]: let $M(t)$ be a martingale with càdlàg paths. Then for integer $p \ge 1$,

$$\mathbb{E}\left[\sup_{0\le s\le t}|M(s)|^p\right]^{1/p} \le C_p\,\mathbb{E}[[M]_t^{p/2}]^{1/p}. \tag{26}$$

Controlling the quadratic variation of the martingale part therefore indirectly bounds the supremum path in any finite interval $[0,t]$.

Proof We find that

$$\left\|X(\hat{t})\right\|_{l,1}^p = \|X(0)\|_{l,1}^p + \int_0^{\hat{t}} F(X(s))\, ds + M(\hat{t}),$$

with F defined in (17) and where the local martingale $M(\hat{t})$ can be defined as in (4), but via integration with *compensated* Poisson processes instead. The quadratic variation can be bounded via Lemma 4.3,

$$\mathbb{E}\left[[M]_{\hat{t}}^{1/2}\right] \leq \mathbb{E}\left[\int_0^{\hat{t}} C(1 + \|X(s)\|_{l,1}^p + \beta_2^V \|X(s)\|_{l,1}^{p+1})\, ds\right]. \tag{27}$$

Assume first that $\beta_2 = 0$. Using the bound in (20) and (21) for the drift part we get

$$\left\|X(\hat{t})\right\|_{l,1}^p \leq \|X(0)\|_{l,1}^p + \int_0^{\hat{t}} C(1 + \|X(s)\|_{l,1}^p)\, ds + |M_{\hat{t}}|.$$

We thus find from Burkholder's inequality and (27) that

$$\mathbb{E}\left[\sup_{s\in[0,\hat{t}]} \|X(s)\|_{l,1}^p\right] \leq \mathbb{E}[\|X(0)\|_{l,1}^p] + \int_0^{\hat{t}} C\left(1 + \mathbb{E}\left[\sup_{s'\in[0,s]} \|X(s')\|_{l,1}^p\right]\right) ds.$$

Upon defining $\|X\|_{l,1}^p(t) := \sup_{s\in[0,t]} \|X(s)\|_{l,1}^p$ this becomes

$$\mathbb{E}[\|X\|_{l,1}^p(\hat{t})] \leq \mathbb{E}[\|X(0)\|_{l,1}^p] + \int_0^t C(1 + E[\|X\|_{l,1}^p(\hat{s})]\, ds.$$

Grönwall's inequality shows that $\mathbb{E}[\|X\|_{l,1}^p(\hat{t})]$ can be bounded in terms of the initial data and time t. The statement of the theorem follows by letting $P \to \infty$ and Fatou's lemma.

Assume finally that $\beta_2 > 0$. We still have the bound (27) which yields

$$\mathbb{E}\left[[M]_{\hat{t}}^{1/2}\right] \leq \int_0^{\hat{t}} C(1 + \mathbb{E}[\|X(s)\|_{l,1}^{p+1}])\, ds \leq (e^{C\hat{t}} - 1)(\mathbb{E}[\|X(0)\|_{l,1}^{p+1}] + 1),$$

by Theorem 4.2. This leaves us with a bound of $\mathbb{E}[\|X\|_{l,1}^p(\hat{t})]$ in terms of initial data $\mathbb{E}[\|X(0)\|_{l,1}^{p+1}]$, and where the previous argument carries through. □

4.3 Continuity

So far we did not discuss the issue of uniqueness of solutions. The reason is that for an integer-valued process, up until an explosion time, each trajectory is uniquely defined by a series of Poisson-distributed events. In this sense uniqueness is an

immediate property in the current setting. However, it is meaningful to prove a much stronger result in the form of a continuity relation over parameter variations. This also serves as a kind of template for doing numerical analysis investigations. We shall follow [17, 18] quite closely (but see also [1, 28]) and we begin by citing the following technical lemma.

Lemma 4.5 ([18, Lemma 2.4], See Also [22, Lemma 2.2]) *Let Π be a unit-rate Poisson process and T_1, T_2 bounded stopping times, all adapted to $\mathcal{F}_t$. Then*

$$\mathbb{E}[|\Pi(T_2)-\Pi(T_1)|] = \mathbb{E}[|T_2-T_1|], \tag{28}$$

$$\begin{aligned}\mathbb{E}[(\Pi(T_2)-\Pi(T_1))^2] &= 2\,\mathbb{E}[|\Pi(T_2)-\Pi(T_1)|(T_1\vee T_2)] \\ &\quad -\mathbb{E}[|T_2^2-T_1^2|]+\mathbb{E}[|T_2-T_1|].\end{aligned} \tag{29}$$

The lemma is interesting in many different types of perturbation bounds. To exemplify we will make a specific investigation as follows. Let $Y(t)$ be defined as in (4) but with all propensities u_r replaced with a perturbed version $u_r^{(\delta)}$. Given a bound

$$\int_0^t u_r(X(s))\,ds \vee \int_0^t u_r^{(\delta)}(Y(s))\,ds \le B(t), \tag{30}$$

as typically achieved under a stopping time, we find from (28) to (29) that

$$\begin{aligned}&\mathbb{E}\left[\left(\Pi_r\left(\int_0^t u_r^{(\delta)}(Y(s))\,ds\right)-\Pi_r\left(\int_0^t u_r(X(s))\,ds\right)\right)^2\right] \\ &\quad\le (2B(t)+1)\,\mathbb{E}\left[\left|\int_0^t u_r^{(\delta)}(Y(s))\,ds-\int_0^t u_r(X(s))\,ds\right|\right],\end{aligned} \tag{31}$$

typically amenable to further bounds. Let us make the formal requirement that the perturbation satisfies

$$|u_r^{(\delta)}(x)-u_r(x)| \le \delta|u_r(x)|. \tag{32}$$

The transport rates obey a simpler linear scaling, and so we similarly let

$$|q_{ijk}^{(\delta)}-q_{ijk}| \le \delta q_{ijk}. \tag{33}$$

We further assume for simplicity that the entire statement of Assumption 4.1 is valid for the perturbed model, and with the same constants. Define the *joint* stopping time

$$\tau_P := \inf_s\{\|X(s)\|^2 \vee \|Y(s)\|^2 > P\}, \text{ and put } \hat{t} := \tau_P \wedge t. \tag{34}$$

To avoid a forest of constants we shall use the notation $A \preceq_C B$, meaning $A \leq CB$ for $C > 0$ some unspecified constant. For example, $\|x\| \preceq_C \|x\|_1$, and also, by (13), in (12), we have $L_r(V_j^{-1}P) \preceq_C L_r(P)$.

We get from (12)

$$\begin{aligned} |u_r^{(\delta)}(y) - u_r(x)| &\leq L_r(P)\|x-y\| + \delta|u_r(y)| \\ &\leq L_r(P)\|x-y\| + \delta(L_r(P)\|y\| + u_r(0)) \leq C_r(P)(\delta + \|x-y\|), \end{aligned} \tag{35}$$

with $C_r(P)$ a P-dependent expression. We can now bound the dynamics of the difference process $(Y - X)(t)$.

Lemma 4.6 *Define $X(t)$ and $Y(t)$ by (4) with $X(0) = Y(0)$ almost surely, and with $Y(t)$ a perturbed trajectory obeying (32)–(33). Then under the joint stopping time $\hat{t}$ in (34),*

$$\mathbb{E}[\|Y(\hat{t}) - X(\hat{t})\|^2] \preceq_C \delta A(P) + B(P) \int_0^t \mathbb{E}[\|Y(\hat{s}) - X(\hat{s})\|]\, ds, \tag{36}$$

where bounds for the P-dependent constants A and B can be found in (37) below.

Proof Let us first bound the difference process for a single species i in a single voxel j. From (4),

$$\begin{aligned} \left(Y_{ij}(\hat{t}) - X_{ij}(\hat{t})\right)^2 = \Bigg(&- \sum_{r=1}^{R} \mathbb{N}_{ri} \left(\Pi_{rj}(\cdot) - \Pi_{rj}(\cdot)\right) \\ &- \sum_{k=1}^{J} \left(\Pi'_{ijk}(\cdot) - \Pi'_{ijk}(\cdot)\right) + \sum_{k=1}^{J} \left(\Pi'_{ikj}(\cdot) - \Pi'_{ikj}(\cdot)\right) \Bigg)^2 \end{aligned}$$

where, for brevity, the local time arguments can be found in (4).

Using the bound on the mesh connectivity (14) and Jensen's inequality we find

$$\left(Y_{ij}(\hat{t}) - X_{ij}(\hat{t})\right)^2 \leq (R + 2M_D)\,(A_1 + A_2 + A_3)\,,$$

where in terms of

$$A_1 = \sum_{r=1}^{R} \mathbb{N}_{ri}^2 \left(\Pi_{rj}\left(\int_0^{\hat{t}} V_j u_r^{(\delta)}(V_j^{-1} Y_{\cdot j}(s))\, ds \right) - \Pi_{rj}\left(\int_0^{\hat{t}} V_j u_r(V_j^{-1} X_{\cdot j}(s))\, ds \right) \right)^2,$$

$$A_2 = \sum_{k=1}^{J} \left(\Pi'_{ijk}\left(\int_0^{\hat{t}} q_{ijk}^{(\delta)} Y_{ij}(s)\, ds \right) - \Pi'_{ijk}\left(\int_0^{\hat{t}} q_{ijk} X_{ij}(s)\, ds \right) \right)^2,$$

$$A_3 = \sum_{k=1}^{J} \left(\Pi'_{ikj}\left(\int_0^{\hat{t}} q_{ikj}^{(\delta)} Y_{ik}(s)\, ds \right) - \Pi'_{ikj}\left(\int_0^{\hat{t}} q_{ikj} X_{ik}(s)\, ds \right) \right)^2.$$

Note that A_2 and A_3 are connected via a simple permutation of the dimensions $j \leftrightarrow k$.

By the Lipschitz assumption (12),

$$\int_0^{\hat{t}} V_j u_r(V_j^{-1} X_{\cdot j}(s))\, ds \le t V_j (L_r(V_j^{-1}P)P + u_r(0)) \le_C L_r(P)P + 1,$$

as well as the identical bound for the perturbed process $Y(t)$. Using this in (30), by Lemma 4.5 (31) we get from (35)

$$\mathbb{E}[A_1] \le_C \sum_r \mathbb{N}_{ri}^2 (L_r(P)P + 1) C_r(P) \left(\delta + \int_0^t \mathbb{E}[\|Y(\hat{s}) - X(\hat{s})\|]\, ds \right).$$

Using similar arguments we readily find

$$\mathbb{E}[A_2] \le_C \sum_k (P+1)P \left(\delta + \int_0^t \mathbb{E}[\|Y(\hat{s}) - X(\hat{s})\|]\, ds \right),$$

and the identical bound for $\mathbb{E}[A_3]$.

Summing over i and j we finally arrive at

$$\begin{aligned} \mathbb{E}\left[\|Y(t) - X(t)\|^2\right] &\le (R + 2M_D) \sum_{i,j} \mathbb{E}\left[A_1 + A_2 + A_3\right] \\ &\le_C \delta A(P) + B(P) \int_0^t \mathbb{E}[\|Y(\hat{s}) - X(\hat{s})\|]\, ds. \end{aligned} \tag{37}$$

□

Theorem 4.7 (Continuity) *Let the two trajectories $X(t)$ and $Y(t)$ obey the assumptions of Lemma 4.6. Then*

$$\lim_{\delta \to 0+} \mathbb{E}[\|Y(t) - X(t)\|^2] = 0. \tag{38}$$

Proof We wish to apply the Grönwall inequality to the inequality of Lemma 4.6. Unfortunately, the exponents on the different sides of (36) differ. One way out is to develop a similar bound as in Lemma 4.6 but for the norm $\|\cdot\|_1$ instead, and then combine the two by using the norm equivalence $\|\cdot\| \le \|\cdot\|_1$. A faster way which applies here is to use the "integer inequality": since both processes are integer valued, the difference process is as well, and hence $\|Y(t) - X(t)\| \le \|Y(t) - X(t)\|^2$. Using this trick and Grönwall's inequality we thus arrive at

$$\mathbb{E}[\|Y(\hat{t}) - X(\hat{t})\|^2] \le_C \delta A(P) \exp(tB(P)). \tag{39}$$

Write

$$\mathbb{E}\left[\|Y(t)-X(t)\|^2\right]=\mathbb{E}\left[\|Y(\hat{t})-X(\hat{t})\|^2\right]+\mathbb{E}\left[\|Y(t)-X(t)\|^2 1_{t>\hat{t}}\right].$$

To bound the remainder, note that by the Cauchy-Schwartz and Markov inequalities,

$$\begin{aligned}\mathbb{E}\left[\|Y(t)-X(t)\|^2 1_{t>\hat{t}}\right] &\leq \left(\mathbb{E}\left[\|Y(t)-X(t)\|^4\right]\right)^{1/2}\left(\mathbf{P}[t\geq\tau_P]\right)^{1/2}\\ &\leq B_1(t)\,P^{-1/2}\,\mathbb{E}\left[\sup_{s\in[0,t]}\|X(s)\|^2\vee\|Y(s)\|^2\right]\\ &\leq B_1(t)\,P^{-1/2}\,B_2(t),\end{aligned}$$

say, where the existence of the two bounds B_1 and B_2 is guaranteed by Theorems 4.2 and 4.4. To conclude,

$$\mathbb{E}\left[\|Y(t)-X(t)\|^2\right]\leq_C \delta A(P)\exp(tB(P))+B_1(t)\,P^{-1/2}\,B_2(t).$$

Given an $\varepsilon>0$ we may first select P large enough that the remainder is $\leq\varepsilon/2$. Subsequently we may pick a δ_0 such that also the stopped part is $\leq\varepsilon/2$ for all $\delta<\delta_0$. This then proves convergence as $\delta\to 0+$. □

It sometimes provides with some insight to look at the bounded case as well. In case we can select P large enough that the remainder vanish altogether, we arrive at the following

Corollary 4.8 (Perturbation Estimate, Bounded Version) *If in Theorem 4.7, the processes $X(t)$ and $Y(t)$ are bounded, then*

$$\mathbb{E}[\|X(t)-Y(t)\|^2]=O(\delta). \tag{40}$$

5 Discussion

In this review we have shown how to develop a few basic stability results for stochastic compartment-based reaction-transport models. Selected analytical techniques have been highlighted and care has been taken in clearly formulating our working assumptions. A strong well-posedness result for a large and relevant class of problems was proved and a continuous dependence on input data was also achieved within the same framework. The latter development is an example of the use of the different parts of the theory in a setting which relates to that of numerical analysis and method's development.

In the bounded setting, not unusually can explicit error bounds be proven. However, a strength of the theory is that it applies in the limit sense also for open systems under the effects of nonlinear feedback terms. One can argue that numerical analysis helps asking the difficult question here: what are the practical boundaries of modeling?

References

1. D.F. Anderson, An efficient finite difference method for parameter sensitivities of continuous time Markov chains. SIAM J. Numer. Anal. **50**(5), 2237–2258 (2012). doi:10.1137/110849079
2. D.F. Anderson, D.J. Higham, Multi-level Monte Carlo for continuous time Markov chains, with applications in biochemical kinetics. Multiscale Model. Simul. **10**(1), 146–179 (2012). doi:10.1137/110840546
3. D.F. Anderson, M. Koyama, Weak error analysis of numerical methods for stochastic models of population processes. Multiscale Model. Simul. **10**(4), 1493–1524 (2012). doi:10.1137/110849699
4. D.F. Anderson, A. Ganguly, T.G. Kurtz, Error analysis of tau-leap simulation methods. Ann. Appl. Probab. **21**(6), 2226–2262 (2011). doi:10.1214/10-AAP756
5. D. Applebaum, *Lévy Processes and Stochastic Calculus*. Cambridge Studies in Advanced Mathematics, vol. 93 (Cambridge University Press, Cambridge, 2004)
6. G. Arampatzis, M. Katsoulakis, P. Plecháč, Parallelization, processor communication and error analysis in lattice kinetic monte carlo. SIAM J. Numer. Anal. **52**(3), 1156–1182 (2014). doi:10.1137/120889459
7. P. Bauer, S. Engblom, Sensitivity estimation and inverse problems in spatial stochastic models of chemical kinetics, in *Numerical Mathematics and Advanced Applications: ENUMATH 2013*, ed. by A. Abdulle, S. Deparis, D. Kressner, F. Nobile, M. Picasso. Lecture Notes in Computational Science and Engineering, vol. 103 (Springer, Berlin, 2015), pp. 519–527. doi:10.1007/978-3-319-10705-9_51
8. P. Bauer, S. Engblom, S. Widgren. Fast event-based epidemiological simulations on national scales. Int. J. High Perform. Comput. Appl. **30**(4), 438–453 (2016). doi:10.1177/1094342016635723
9. E. Blanc, S. Engblom, A. Hellander, P. Lötstedt, Mesoscopic modeling of stochastic reaction-diffusion kinetics in the subdiffusive regime. Multiscale Model. Simul. **14**(2), 668–707 (2016). doi:10.1137/15M1013110
10. P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Texts in Applied Mathematics, vol. 31 (Springer, New York, 1999)
11. C. Briat, A. Gupta, M. Khammash, A scalable computational framework for establishing long-term behavior of stochastic reaction networks. PLoS Comput. Biol. **10**(6), e1003669 (2014). doi:10.1371/journal.pcbi.1003669
12. A. Chevallier, S. Engblom, Pathwise error bounds in multiscale variable splitting methods for spatial stochastic kinetics, 2016. Available at https://arxiv.org/abs/1607.00805
13. M. Dobrzyński, J.V. Rodríguez, J.A. Kaandorp, J.G. Blom, Computational methods for diffusion-influenced biochemical reactions. Bioinformatics **23**(15), 1969–1977 (2007). doi:10.1093/bioinformatics/btm278
14. B. Drawert, S. Engblom, A. Hellander, URDME: a modular framework for stochastic simulation of reaction-transport processes in complex geometries. BMC Syst. Biol. **6**(76), 1–17 (2012) doi:10.1186/1752-0509-6-76
15. S. Engblom, Parallel in time simulation of multiscale stochastic chemical kinetics. Multiscale Model. Simul. **8**(1), 46–68 (2009). doi:10.1137/080733723

16. S. Engblom (ed.), *Student's Book: Numerical Functional Analysis* (Uppsala University, Uppsala, 2014). Available at http://www.it.uu.se/grad/courses/scicomp/NumFunkAnalysis/NFAStudentBook.pdf
17. S. Engblom, On the stability of stochastic jump kinetics. Appl. Math. **5**(19), 3217–3239 (2014) doi:10.4236/am.2014.519300
18. S. Engblom, Strong convergence for split-step methods in stochastic jump kinetics. SIAM J. Numer. Anal. **53**(6), 2655–2676 (2015). doi:10.1137/141000841
19. S. Engblom, L. Ferm, A. Hellander, P. Lötstedt, Simulation of stochastic reaction-diffusion processes on unstructured meshes. SIAM J. Sci. Comput. **31**(3), 1774–1797 (2009). doi:10.1137/080721388
20. S.N. Ethier, T.G. Kurtz, *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics (Wiley, New York, 1986)
21. D. Fange, J. Elf, Noise-induced Min phenotypes in *E. coli*. PLoS Comput. Biol. **2**(6), 637–648 (2006). doi:10.1371/journal.pcbi.0020080
22. A. Ganguly, D. Altıntan, H. Koeppl, Jump-diffusion approximation of stochastic reaction dynamics: error bounds and algorithms. Multiscale Model. Simul. **13**(4), 1390–1419 (2015). doi:10.1137/140983471
23. C.W. Gardiner, *Handbook of Stochastic Methods*. Springer Series in Synergetics, 3rd edn. (Springer, Berlin, 2004)
24. D.T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. **22**(4), 403–434 (1976). doi:10.1016/0021-9991(76)90041-3
25. E. Hausenblas, SPDEs driven by Poisson random measure with non Lipschitz coefficients: existence results. Probab. Theory Relat. Fields **137**(1–2), 161–200 (2007). doi:10.1007/s00440-006-0501-8
26. D.J. Higham, P.E. Kloeden, Numerical methods for nonlinear stochastic differential equations with jumps. Numer. Math. **101**(1), 101–119 (2005). doi:10.1007/s00211-005-0611-8
27. J. Karlsson, R. Tempone, Towards automatic global error control: computable weak error expansion for the tau-leap method. Monte Carlo Methods Appl. **17**(3), 233–278 (2011). doi:10.1515/MCMA.2011.011
28. T.G. Kurtz, Representation and approximation of counting processes, in *Advances in Filtering and Optimal Stochastic Control*, ed. by W.H. Fleming, L.G. Gorostiza. Lecture Notes in Control and Information Sciences, vol. 42, pp. 177–191 (Springer, Berlin, 1982). doi:10.1007/BFb0004537
29. P.D. Lax, R.D. Richtmyer, Survey of the stability of linear finite difference equations. Commun. Pure Appl. Anal. **9**(2), 267–293 (1956). doi:10.1002/cpa.3160090206
30. T. Li, Analysis of explicit tau-leaping schemes for simulating chemically reacting systems. Multiscale Model. Simul. **6**(2), 417–436 (2007). doi:10.1137/06066792X
31. S.P. Meyn, R.L. Tweedie, Stability of Markovian processes III: Foster-Lyapunov criteria for continuous-time processes. Adv. Appl. Probab. **25**(3), 518–548 (1993)
32. B. Øksendal, T. Zhang, The Itô-Ventzell formula and forward stochastic differential equations driven by Poisson random measures. Osaka J. Math. **44**(1), 207–230 (2007)
33. P.E. Protter, *Stochastic Integration and Differential Equations*. number 21 in Stochastic Modelling and Applied Probability, 2nd edn. (Springer, Berlin, 2005)
34. M. Rathinam, Moment growth bounds on continuous time Markov processes on non-negative integer lattices. Quart. Appl. Math. **73**(2), 347–364 (2015). doi:10.1090/S0033-569X-2015-01372-7
35. R. Situ, *Theory of Stochastic Differential Equations with Jumps and Applications*. Mathematical and Analytical Techniques with Applications to Engineering (Springer, New York, 2005)
36. N.G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 2nd edn. (Elsevier, Amsterdam, 2004)
37. C.A. Yates, M.B. Flegg, The pseudo-compartment method for coupling partial differential equation and compartment-based models of diffusion. J. R. Soc. Interface **12**(106) (2015). doi:10.1098/rsif.2015.0141

The T Cells in an Ageing Virtual Mouse

Mario Castro, Grant Lythe, and Carmen Molina-París

1 Introduction

The multiscale problem that a modeller in biology is presented with, trying to provide a systematic description of many agents, their properties, their internal dynamics and interactions, is daunting. On the other hand, biology provides a natural scale, with individual cells as agents. In agent-based computation, variables representing cell population sizes may be evaluated by counting cells of various types, but the governing dynamical rules are laid down one event at a time [1, 2]. Every cell is an individual, with its own set of attributes (state of activation, surface molecule profile, spatial location, for example). Populations of cells decrease or increase because individual cells die or divide. Here, by way of a tutorial on agent-based immune system modelling, we implement a model of the behaviour of the set of T cells in a body—numbering more than 10^{11} in an adult human, and more than 10^7 in an adult mouse [3].

Each T cell is characterised by the T-cell receptors (TCRs) found on its surface; it either began life in the thymus, or is descended from a cell, with the same TCR, that began life in the thymus and exited to circulate through the rest of the body (the periphery). The number of T cells in the periphery increases when new T cells exit the thymus or when an existing cell divides into two daughter cells, and decreases when an existing cell dies. Our adaptive immune system relies on the enormous

M. Castro
Department of Applied Mathematics, University of Leeds, Leeds, UK

Grupo Intedisciplinar de Sistemas Complejos (GISC), Universidad Pontificia Comillas, Madrid, Spain

G. Lythe (✉) • C. Molina-París
Department of Applied Mathematics, University of Leeds, Leeds, UK
e-mail: grant@maths.leeds.ac.uk

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_6

diversity of the TCR repertoire [4], which can be thought of as an ecological system, the set of cells with the same TCR being a species, or clonotype, that competes with many others. The response to an infection is the average of that of a small number of clonotypes that expand greatly and many clonotypes that do not [5, 6]; even T cells bearing identical TCRs undergo heterogeneous proliferation and development [7, 8].

Every T cell in our model is either a $CD4^+$ or a $CD8^+$ T cell. This classification is based on the observed predominance of one of the two co-receptor surface molecules, CD4 or CD8, on any one T cell. In the immune system's response to infection, the primary role of $CD4^+$ T cells is to produce chemical signals that "help" other cells of the immune system, while that of $CD8^+$ T cells is to produce perforin and granzymes that are toxic to cancerous, infected or damaged cells [9]. Models that distinguish "naive" T cells from regulatory T cells and from different types of memory T cells, based on surface markers such as CCR7 [10], and track population subsets in different organs of the body [11], are not considered here.

Construction of a mathematical model of the in-host kinetics of an infection usually begins by enumerating the relevant cell populations. Each population may be represented as one real variable in a set of ordinary differential equations whose steady states can be found and that can be solved numerically. For example, influential mathematical models of HIV infection consider four populations: uninfected T cells, latently infected T cells, actively infected T cells, and free virus [12]. Experimental research on infectious disease and the immune system, in recent decades, has revealed more and more cell types. Models that have one variable for the average size of each relevant population can be devised and solved numerically, but are often difficult to interpret. Somewhat more realistic are stochastic models, which can describe fluctuating populations and experimental variability [13]. An important advantage of stochastic models in which population sizes are integers is that the phenomenon of extinction of populations is included in a natural way [14]. However, whether a model is deterministic or stochastic, treating all cells within a population as identical or statistically identical comes at the cost of ignoring cell-to-cell heterogeneity and the receptor–ligand interactions at particular cell–cell interfaces [15–17]. All these features of biological systems are characteristic of agent-based models, in which the fundamental objects are not populations of cells, based on an a priori classification, but individual cells.

Despite the increasing popularity of agent-based modelling in many fields, there is no standard implementation. Some de-facto standards, like Netlogo [18], Repast [19] or SPARK [20], appear in the Systems Biology literature [21]. We find that open-source language `python` provides a simple yet powerful tool to implement, distribute and maintain computational models. In this chapter, we will build, step by step, object-oriented codes exhibiting the main procedures and strategies.

2 Writing a Cell-Based Code

2.1 A Cell is Born

Our agent-based model is implemented in an object-oriented code, that begins by defining the class of objects. In `python`, providing such a template is straightforward. The code shown in Fig. 1 defines the T-cell class, the attribute `a` representing activation, and a method for activating a cell (changing the attribute from `False` to `True`).

2.2 Create a List of Cells

Of course, we are typically interested in a population of more than one cell. In the code shown in Fig. 2, multiple cells are created as instances of the class `T` and stored as elements of the array `Celllist`. In this simple example, where each cell has only one attribute that takes the value `True` or `False`, a population-based model can be constructed, with one population for each value of the attribute, that gives a concise description of the time evolution. However, because each cell is stored separately in the computer's memory, it is easy to envisage the heterogeneous population that is created when there are more cell attributes and some attributes are real numbers.

```
# Tcell.py   GDL and MC 2017
class Cell(object):
    '''generic cell class'''
    def __str__(self): # This method formats the output of 'print' calls
        infostring="Cell"
        return infostring

class T(Cell): # Inherit from Cell class
    """T cell class.
    Attribute: a, activation state (initially False)"""
    def __init__(self):
        Cell.__init__(self) # Call to parent constructor
        self.a = False # Is activated? (not initially)
    def activate(self):
         self.a = True # Activate the cell

myfirstcell = T()
print 'myfirstcell is a',myfirstcell
print 'myfirstcell is activated:',myfirstcell.a
myfirstcell.activate() # Call to method activate
print 'myfirstcell is activated:',myfirstcell.a # now self.a is True
```

Fig. 1 Define a T-cell class. Create a T cell and activate it

```
# Tcells.py   GDL and MC 2017
import random
class Cell(object):
    '''generic cell class'''
    def __str__(self):
        infostring ="Cell"
        return infostring

class T(Cell): # Inherit from class Cell
    """T cell class.
    Attributes: a, activation state (initially False)"""
    def __init__(self):
        Cell.__init__(self) # Call to parent constructor
        self.a = False  # Initially, the cell is not activated
    def activate(self):
         self.a = True

ncells = 10
Celllist = [T() for i in range(ncells)] # Create T for every i in 0,...,ncells -1
print 'a before:',[tcell.a for tcell in Celllist] # Print 'a' for every tcell in Celllist

firstcell = Celllist[0] # Alias for the first T cell
anothercell = random.choice(Celllist) # Pick a random T cell from Celllist
firstcell.activate() # Activate the first cell
anothercell.activate() # Activate a random cell

print 'a after:',[tcell.a for tcell in Celllist] # Print 'a' for every tcell in Celllist
```

Fig. 2 Create ten T cells. Activate the first of them, and one other

2.3 The Scheduler: Birth and Death of Cells

Now let us introduce the fundamental events of cell death and cell division. In the code shown in Fig. 3, each of the cells in our population, independently, may die or divide with probability 0.5.

In Fig. 4 we summarise the steps undertaken in the modelling process.

2.4 The Scheduler: A Gillespie Algorithm Code

In the thymus, developing cells produce new TCRs by a process of gene rearrangement and express, for a time, both the CD4 and CD8 co-receptor molecules. A cell that survives "positive" and "negative" selection" [22, 23] emerges expressing, predominantly, only one of the molecules. It remains so for the rest of its life and passes on these characteristics to its progeny, both in the thymus, when it is classified as an SP4 or an SP8 thymocyte [24–26], and in the rest of the body, when it is classified as a $CD4^+$ or a $CD8^+$ T cell.

In the following code, two subclasses are defined: $CD4^+$ and $CD8^+$ T cells. The dynamics of a cell population are simulated, one event at a time, using the Gillespie algorithm [27, 28]. There are only two types of events, death and division. Thus, there are twice as many candidates for the next event at any time as there are cells at that time (Fig. 5).

```
# bdTcells.py   GDL and MC 2017
from random import randrange,random
import copy

class Cell(object):
   '''setup so that cell types inherit a counter'''
   number = 0
   def __init__(self):
      type(self).number += 1 # Add 1 to the class counter
   def __del__(self):
      type(self).number -= 1 # Decrease the class counter

   def celldivision(self):
       '''create two identical cells from one'''
       newcell = copy.deepcopy(self) # Make a 'clone' of the cell bit by bit
       type(self).number += 1 # Increase the counter
       return newcell

class T(Cell):
   '''T cell class'''
   def __init__(self):
      Cell.__init__(self)
      self.a = False
   def activate(self):
       self.a = True

ncells = 10
Celllist = [T() for i in range(ncells)] # Create T for every i in 0...ncells-1
print 'Start with:',T.number,'T cells'

for cell in Celllist:
   urv=random() # Uniform random number
   if urv < 0.5: # 'Coin flip'. Heads: kill Tails: clone
      Celllist.remove(cell) # Remove element cell from Celllist
      del cell
   else:
      newcell = Cell.celldivision(cell)
      Celllist.append(newcell)

   assert len(Celllist)==T.number,'How many cells?' # Check if we are counting properly

print 'Now there are:',T.number,'T cells'
```

Fig. 3 Flip a coin to decide whether each cell dies or divides

3 Case Study: The T Cells in a Mouse from Infancy to Old Age

The total number of T cells in a mouse is in the range 10^7–10^8. Over the lifetime of a mouse, new cells are produced by the thymus, cells may divide in the periphery, and cells die. Is it feasible to recreate this on a computer?

We construct our code based on the measurements of Hogan et al. [26]. In the thymii of mice of different ages, they counted SP4 and SP8 cells. Because they are in the last stage of thymic development, ready for export to the periphery, their numbers serve as a proxy for the rate of thymic production and the ratio of $CD4^+$ to $CD8^+$ cells among thymic emigrants. With considerable variation from mouse to mouse, an overall exponential decline in rate with half life of about 150 days, and SP4:SP8 ratio of about 4, are estimated from the data in Fig. 6.

Hogan et al. estimate the mean rate of division of cells in the periphery by taking a sample of cells and measuring the fraction of $CD4^+$ and $CD8^+$ T cells,

denoted Ki67$^+$ cells, that display sufficient quantities of the surface molecule Ki67. Suppose that each daughter T cell is Ki67$^+$ for a total of 4 days during and after the time spent cycling before division and that a T cell has a probability λ per day of dividing (into two Ki67$^+$ daughter cells). Then the fraction of Ki67$^+$ cells at any time is twice the number of cells that have divided in the 4 days up to that time: $2 \times 4\,\text{days} \times \lambda$. Based on the observation that 4% of naive cells are Ki67$^+$ [26], $\lambda = 0.04/8 = 1/200\,\text{day}^{-1}$, so the mean time between divisions of peripheral naive CD4$^+$ or CD8$^+$ T cells is 200 days.

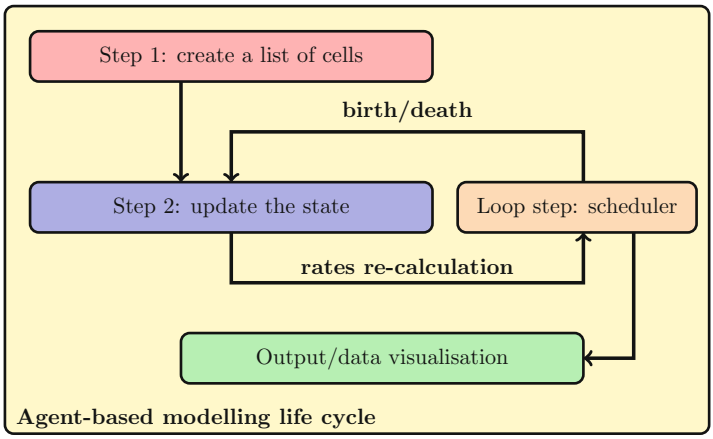


Fig. 4 An illustration of the structure of the code

```
# Gillespie.py   GDL and MC 2017
from random import randrange,random,choice
from numpy import cumsum,log,searchsorted,cumsum
import pylab,copy

class Cell(object):
   ''' setup so that cell types inherit a counter '''
   number = 0
   def __init__(self):
      type(self).number += 1
   def __del__(self):
      type(self).number -= 1

class T(Cell):
   ''' T cell class '''
   def __init__(self):
      Cell.__init__(self)
      self.a = False

class CD4(T):
   ''' CD4 T cell class '''
   def __init__(self):
      T.__init__(self)

class CD8(T):
   ''' CD8 T cell class '''
```

Fig. 5 Two populations of T cells

```
    def __init__(self):
        T.__init__(self)

def death(thistype):
    ''' a cell dies '''
    thislist = celllists[thistype]
    thiscell = thislist.pop(randrange(len(thislist)))
    del thiscell

def birth(thistype):
    ''' a cell divides '''
    thislist = celllists[thistype]
    thislist.append(celltypes[thistype]())

def makerates():
    ''' construct a list of rates of the Gillespie step '''
    rates = []
    ntotal = CD4.number + CD8.number
    for celltype in celltypes:
        rates.append(mu*celltypes[celltype].number)
        rates.append(gamma*celltypes[celltype].number/ntotal)
    return rates,sum(rates)

mu,gamma,ncells = 0.1,100.0,10 # Model parameters
CD4list = [CD4() for i in range(ncells)] # List of CD4s
CD8list = [CD8() for i in range(ncells)] # List of CD8s

cellnames=['CD4','CD8']
celllists={'CD4':CD4list,'CD8':CD8list} # Dictionary of cell lists
celltypes={'CD4':CD4,'CD8':CD8} # Dictionary of cell types

events={0:death,1:birth,2:death,3:birth} # Possible Gillespie events

t,tmax = 0.0,100.0 # Initial and final simulation time

tforplot,cd4forplot,cd8forplot = [t],[CD4.number],[CD8.number] # Auxiliary lists
while t < tmax:
    rates,sumrates = makerates() # Create rates from current state
    i = searchsorted(cumsum(rates),random()*sumrates) # Pick one event randomly
    events.get(i)(cellnames[i/2]) # Use dictionaries to call the right event
    t -= log(random())/sumrates # Increment time
    tforplot.append(t) # Add current time for plotting
    cd4forplot.append(CD4.number)
    cd8forplot.append(CD8.number)
print 'At t =',t,'there are',CD4.number,'CD4 cells and',CD8.number,'CD8 cells' # Output

# The following 6 lines create the plot
pylab.step(tforplot,cd4forplot,label='CD4 cells')
pylab.step(tforplot,cd8forplot,label='CD8 cells')
pylab.xlabel('$t$')
pylab.ylabel('number of cells')
pylab.legend(loc='best')
pylab.show()
```

Fig. 5 (continued)

Based on the rates estimated in Hogan et al. [26] and den Braber et al. [29], we take the daily production of the thymus of a t-day-old mouse to be

$$\theta(t) = 10^6 \exp(-\nu(t - 56)) \tag{1}$$

cells, where $\nu = 0.004\,\text{day}^{-1}$. The fraction of these cells that are $CD4^+$ is 80%. Based on the same sources, we take the death rates to be $0.030\,\text{day}^{-1}$ for $CD4^+$ cells and $0.015\,\text{day}^{-1}$ for $CD8^+$ cells. As can be seen in Fig. 7, even if the number of $CD4^+$ cells exiting the thymus is four times larger than the number of $CD8^+$ cells exiting the thymus, it is possible that the number of peripheral $CD8^+$ T cells approaches or even overtakes that of $CD4^+$ T cells in late adulthood.

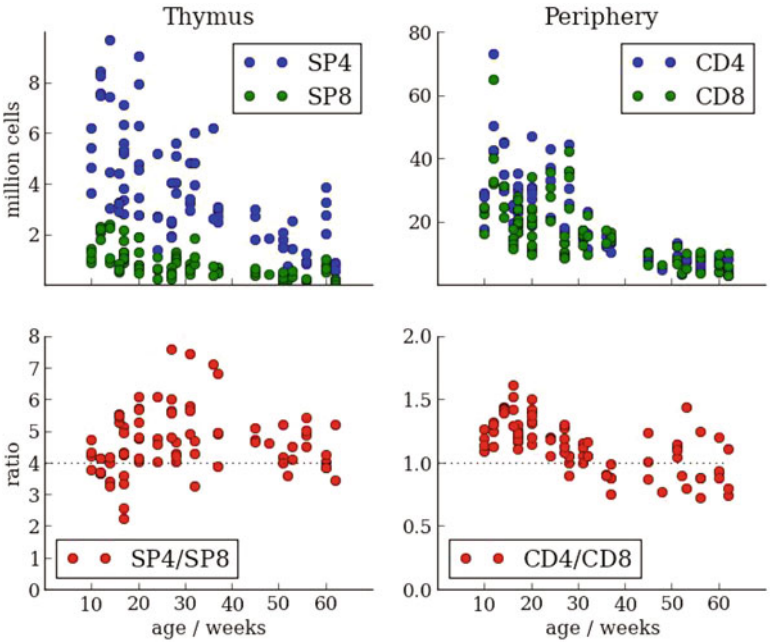


Fig. 6 Cell counts from mice thymii (*left*) and periphery (lymph nodes plus spleen, *right*). Data, from 82 mice, kindly provided by Thea Hogan

3.1 Heterogeneous Populations: Every Cell is Different

How can the Ki67 marker be incorporated into the computational model? The simplest idea would be to give the T cell class a suitable attribute, for example:

```
class T(Cell):
    ''' T cell class '''
    def _ _init_ _(self):
        Cell._ _init_ _(self)
        self.ki67 = 'lo'
    def upk(self):
         self.ki67 = 'hi'
    def downk(self):
         self.ki67 = 'lo'
```

If the method `upk` is called on cell division, `ki67` attribute is set to `hi` on division. It will remain so for an exponentially distributed time with an average of four days if an extra type of event, calling the method `downk` with rate 0.25 day^{-1} per cell, is defined. More realistic models of cell dynamics can be defined by using non-exponential distributions, which is achieved by giving a new attribute, the down-regulation time, to each daughter cell. Models that better reproduce

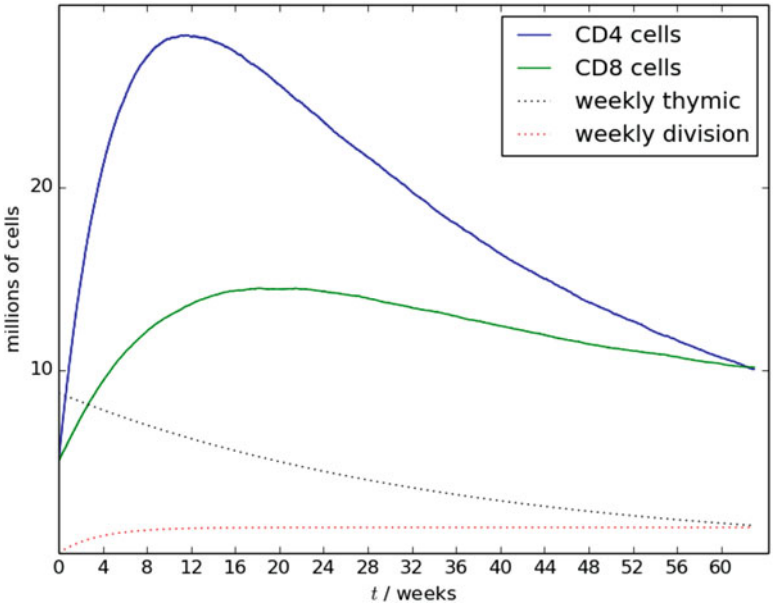


Fig. 7 The number of $CD4^+$ cells and $CD8^+$ cells as a function of the age of the mouse, based on the assumptions described in Sect. 3. The *dotted lines* are the total numbers of cells, per week, exiting the thymus, given by (1), and the total number of cell divisions per week. The graph was produced by the code in Appendix

cellular heterogeneity and mimic the experimental methodology of classification are produced if each cell's `ki67` attribute is a real number, corresponding to the fluorescence intensity used in flow cytometry as a cell-by-cell measure of Ki67. Then, in silico as in the laboratory, whether a cell is deemed to be $Ki67^+$ or not depends on a user-defined threshold value.

In the version of the code given in Appendix, each T cell is given an attribute `tcr` which is an integer label of its TCR. New cells emerging from the thymus are given a new value at random, cells that divide in the periphery pass on their `tcr` value to their daughters. The code reproduced in Appendix, and available at https://github.com/mariocastro73/CellCulture/blob/master/MouseChapter.py, runs on a desktop machine in a few hours. The run time and maximum memory requirements depend on the value chosen for the scaling factor `sfac`. In Fig. 7, we used `sfac`$= 0.4$.

4 Discussion

The agent-based code provided in Appendix is able to follow all the birth and death events that happen in a number of cells similar to the number of T cells in a mouse (about 10^7 in adulthood) over the lifetime of a mouse. The main obstacle, on our

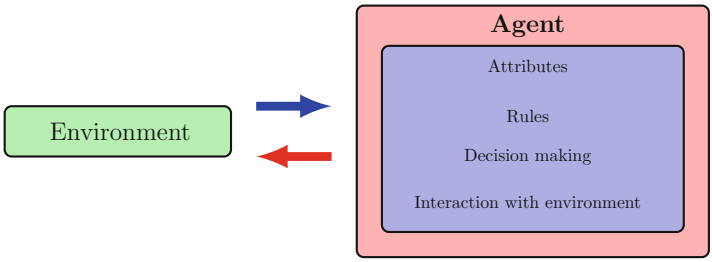


Fig. 8 Agent-based model scheme

desks, to scaling up to numbers of T cells in humans [30–32], and a set of attributes that will capture the complexities of memory and regulatory T cells, is memory storage rather than computation time.

From a practical perspective, agent-based modelling has two benefits. On the one hand, it accommodates cell-level heterogeneity that cannot be implemented in population-based descriptions. On the other hand, we can easily track the fate of individual cells and classify them into subpopulations according to a prescribed criterion, opening the door to coarse-grained descriptions, that are, in some cases, amenable to analytical treatments.

We close with the definition of agent-based modelling provided by Nigel Gilbert [33]:

> *Agent-based modelling is a computational method that enables a researcher to create, analyse, and experiment with models composed of agents that interact within an environment.*

This idea is illustrated in Fig. 8.

Acknowledgements We are grateful for discussions with, and data provided by, Thea Hogan, Ben Seddon and Andy Yates. GDL thanks the Isaac Newton Institute programme *Stochastic Dynamical Systems in Biology: Numerical Methods and Applications*. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 317040 (QuanTI). This work has been partially supported by grants FIS2013-47949-C2-2-P (Spain), PIRSES-GA-2012-317893 (7th FP, EU), and BIOCAPS (FP7/REGPOT-2012- 2013.1, EC) under grant agreement no. 316265. MC thanks the Salvador de Madariaga programme through grant PRX16/00287.

Appendix: Mouse T Cell Repertoire Lifetime Code

```
# Mouse T-cell repertoire model.  GL and MC 2017
# Thymic production plus death and division in the periphery.
# Age-dependence of thymus and body mass from Hogan et al. PNAS (2015)
# Time is measured in days. Gillespie algorithm
from __future__ import print_function  # use python3 notation print()
from random import randrange,random,choice
from numpy import log,searchsorted,array
from math import exp
import matplotlib
matplotlib.use('Agg')
import pylab,copy,datetime

''' Model of the T-cells in a mouse '''

class Cell(object):
    ''' setup so that cell types inherit a counter '''
    number = 0
    def __init__(self):
        type(self).number += 1 # Increase class counter for every new cell
    def __del__(self):
        type(self).number -= 1 # Decrease class counter for every death

class T(Cell):
    ''' T cell class '''
    def __init__(self,tcr=None):
        Cell.__init__(self)
        if tcr == None:
            self.tcr = randrange(0,1e12) # Possible different clonotypes
        else:
            self.tcr = tcr # If passed in the constructor, it's inherited

class CD4(T):
    ''' CD4 T cell class '''
    def __init__(self,tcr=None):
        T.__init__(self,tcr)

class CD8(T):
    ''' CD8 T cell class '''
    def __init__(self,tcr=None):
        T.__init__(self,tcr)

class CellPopulation(object):
    ''' Collection of cells and methods to manipulate them '''
    def __init__(self,ncells):
        self.cellnames=['CD4','CD8']
        self.celltypes={'CD4':CD4,'CD8':CD8}
        ###### Initial list of cells in the population #####
        self.CD4list = [CD4() for i in range(ncells//2)]
        self.CD8list = [CD8() for i in range(ncells//2)]
        self.celllists={'CD4':self.CD4list,'CD8':self.CD8list}
        ###### Methods to manipulate cells #####
        self.events={0:self.death, 1:self.division, 2:self.thymus}

    def death(self,thistype):
        ''' a cell dies '''
        index = randrange(0,self.celltypes[thistype].number) # Randomly chosen cell
        thiscell = self.celllists[thistype][index]
        del self.celllists[thistype][index]

    def division(self,thistype):
        ''' a cell divides '''
        index = randrange(0,self.celltypes[thistype].number) # Randomly chosen cell
        tcr = self.celllists[thistype][index].tcr
        newcell = self.celltypes[thistype](tcr)
        self.celllists[thistype].append(newcell)

    def thymus(self,thistype):
        ''' creation of cells of a new clonotype, with nthy cells '''
        thislist = self.celllists[thistype]
        extendlist = [self.celltypes[thistype]() for i in range(nthy)]
        thislist.extend(extendlist)
```

(continued)

```
72
73 class simulation(object):
74     ''' Simulation class '''
75     def __init__(self,tmax):
76         self.tmax = tmax # Maximum simulation time
77         self.t = 0 # Age of the mouse. Starting at 0
78         self.tforplot = [self.t] # Auxiliary variables for output
79         self.cd4forplot = [CD4.number] # Auxiliary variables for output
80         self.cd8forplot = [CD8.number] # Auxiliary variables for output
81         self.ratioforplot = [f8] # Auxiliary variables for output
82         self.cellpop = CellPopulation(ncells) # Create ncells
83
84     def scheduler(self):
85         ''' Scheduler '''
86         while self.t < self.tmax: # While not at tmax
87             self.gamma = gmax*(1-exp(-10*nu*self.t)) # peripheral division
88             self.thyout = thymax*exp(-nu*(self.t-56))# thymic production
89             self.theta = {'CD4':self.thyout*(1-f8)/nthy,'CD8':self.thyout*f8/nthy}
90             rates,sumrates = self.makerates() # Create gillespie rates
91             i = searchsorted(rates,random()*sumrates) # Find the chosen event
92             self.cellpop.events.get(i%3)(self.cellpop.cellnames[i//3]) # Execute one event
93             self.t -= log(random())/sumrates # Gillespie time update
94
95             if int(self.t*24) != int(self.tforplot[-1]*24): # Store for output
96                 self.tforplot.append(self.t)
97                 self.cd4forplot.append(CD4.number/sfac)
98                 self.cd8forplot.append(CD8.number/sfac)
99                 self.ratioforplot.append(1.*CD4.number/CD8.number)
100                if int(self.tforplot[-1]/7) != int(self.tforplot[-2]/7):
101                    print('At',int(self.t/7),'weeks, there are',CD4.number,
102                          'CD4 cells and',CD8.number,'CD8 cells.')
103                    print('Daily thymic production is',int(self.thyout),
104                          'and peripheral division',int(self.gamma))
105
106    def makerates(self):
107        ''' construct a list of rates of the Gillespie step '''
108        rates = []
109        ntotal = CD4.number + CD8.number
110        aux=0
111        for celltype in self.cellpop.celltypes:
112            aux = aux + mu[celltype]*self.cellpop.celltypes[celltype].number # Death
113            rates.append(aux)
114            aux = aux + self.gamma*self.cellpop.celltypes[celltype].number/ntotal # Division
115            rates.append(aux)
116            aux = aux + self.theta[celltype] # Create new clonotypes in the thymus
117            rates.append(aux)
118        return rates,aux
119
120    def visualization(self):
121        ''' create a plot with the time course of CD4's and CD8's '''
122        pylab.clf() # Clear the previous plot
123        pylab.step(self.tforplot,self.cd4forplot,label='CD4 cells')
124        pylab.step(self.tforplot,self.cd8forplot,label='CD8 cells')
125        pylab.plot(self.tforplot,[7*thymax*exp(-nu*(t-56))/sfac for t in self.tforplot],
126                   ':k',label='weekly thymic')
127        pylab.plot(self.tforplot,[7*gmax*(1-exp(-10*nu*t))/sfac for t in self.tforplot],
128                   ':r',label='weekly division')
129        pylab.xlabel('$t$ / weeks')
130        pylab.xticks([28*i for i in range(int(tmax/28)+1)],
131                     [4*i for i in range(int(tmax/14)+1)])
132        pylab.ylabel('millions of cells')
133        pylab.yticks([1e7,2e7],[10,20])
134        pylab.legend(loc='upper right')
135        mydate = datetime.datetime.today().strftime("%d%b%H%M")
136        pylab.savefig('Mouse-'+str(int(100*sfac))+str(mydate)+'.png')
137        pylab.clf() # Clear the previous plot
138        pylab.step(self.tforplot,self.ratioforplot,label='CD4/CD8 ratio')
139        pylab.plot(self.tforplot,[1 for t in self.tforplot])
140        pylab.xlabel('$t$ / weeks')
141        pylab.xlim([0,max(self.tforplot)])
142        pylab.xticks([28*i for i in range(int(tmax/28)+1)],
```

(continued)

```
                    [4*i for i in range(int(tmax/14)+1)])
        pylab.ylim([0,max(self.ratioforplot)*1.1])
        pylab.ylabel('ratio CD4 to CD8')
        pylab.legend(loc='upper right')
        mydate = datetime.datetime.today().strftime("%d%b%H%M")
        pylab.savefig('Mouse-ratio'+str(int(100*sfac))+str(mydate)+'.png')

#################### global parameter values ##################################
# scaling factor
sfac = 0.10  # fraction of the whole mouse (use sfac = 1 for whole mouse)
# thymic production:
nthy,f8 = 4,1.0/5  # cells per new clone, fraction CD8
thymax,nu = 1000000*sfac,0.004 # daily thymic rate at 8 weeks, decay rate
# periphery:
gmax,mu = 200000*sfac,{'CD4':0.030,'CD8':0.015}  # peripheral division and death
ncells = int(10*thymax) # initial cell number
tmax = 63*7 # Total number of days of the mouse's life
###############################################################################

sim = simulation(tmax) # Create a new simulation
sim.scheduler() # Start the scheduler
sim.visualization() # At the end of the simulation, produce the output

###############################################################################
# python2.7 is recommended
# use // for integer division, works in python2 and python3
###############################################################################
```

References

1. J.L. Segovia-Juarez, S. Ganguli, D. Kirschner, Identifying control mechanisms of granuloma formation during *M. tuberculosis* infection using an agent-based model. J. Theor. Biol. **231**(3), 357–376 (2004)
2. J. Cosgrove, J. Butler, K. Alden, M. Read, V. Kumar, L. Cucurull-Sanchez, J. Timmis, M Coles, Agent-based modeling in systems pharmacology. CPT: Pharmacometrics Syst. Pharmacol. **4**(11), 615–629 (2015)
3. M.K. Jenkins, H.H. Chu, J.B. McLachlan, J.J. Moon, On the composition of the preimmune repertoire of T cells specific for peptide-major histocompatibility complex ligands. Ann. Rev. Immunol. **28**, 275–294 (2010)
4. A.W. Goldrath, M.J. Bevan, Selecting and maintaining a diverse T cell repertoire. Nature **402**, 255–262 (1999)
5. P.D. Hodgkin, Concepts for the development of a quantitative theory of clonal selection and class regulation using lessons from the original. Immunol. Cell Biol. **86**(2), 161–165 (2008)
6. F.M. Burnet, A modification of Jerne's theory of antibody production using the concept of clonal selection. CA Cancer J. Clin. **26**(2), 119–121(1976)
7. E.D. Hawkins, J.F. Markham, L.P. McGuinness, P.D. Hodgkin, single-cell pedigree analysis of alternative stochastic lymphocyte fates. Proc. Natl. Acad. Sci. **106**(32), 13457–13462 (2009)
8. C. Gerlach, J.C. Rohr, L. Perié, N. van Rooij, J.W.J. van Heijst, A. Velds, J. Urbanus, S.H. Naik, H. Jacobs, J.B. Beltman et al., Heterogeneous differentiation patterns of individual $CD8^+$ T cells. Science **340**(6132), 635–639 (2013)
9. A. Ahmed, D. Nandi, T cell activation and function: role of signal strength, in *Mathematical Models and Immune Cell Biology*, ed. by C. Molina-París, G. Lythe (Springer, London, 2011), pp. 77–100
10. F. Sallusto, D. Lenig, R. Förster, M. Lipp, A. Lanzavecchia, Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. Nature **402**, 34–38 (1999)
11. J.J.C. Thome, B. Grinshpun, B.V. Kumar, M. Kubota, Y. Ohmura, H. Lerner, G.D. Sempowski, Y. Shen, D.L. Farber, Long-term maintenance of human naïve t cells through in situ homeostasis in lymphoid tissue sites. Sci. Immunol. **1**(6), eaah6506 (2016)

12. A.S. Perelson, D.E. Kirschner, R. De Boer, Dynamics of HIV infection of $CD4^+$ T cells. Math. Biosci. **114**(1), 81–125 (1993)
13. S.M. Ciupe, B.H. Devlin, M.L. Markert, T.B. Kepler, The dynamics of T-cell receptor repertoire diversity following thymus transplantation for digeorge anomaly. PLoS Comput. Biol. **5**(6), 1–13 (2009)
14. G. Lythe, R.E. Callard, R.L. Hoare, C. Molina-París, How many TCR clonotypes does a body maintain? J. Theor. Biol. **389**, 214–224 (2016)
15. R. Varma, TCR triggering by the pMHC complex: valency, affinity, and dynamics. Sci. STKE **1**(19), pe21 (2008)
16. M.S. Kuhns, M.M. Davis, TCR signaling emerges from the sum of many parts. Front. Immunol. **3**, 1–13 (2012)
17. J.F. Allard, O. Dushek, D. Coombs, P. Anton van der Merwe, Mechanical modulation of receptor–ligand interactions at cell–cell interfaces. Biophys. J. **102**(6), 1265–1273 (2012)
18. S. Tisue, U. Wilensky, Netlogo: a simple environment for modeling complexity, in *International Conference on Complex Systems*, Boston, MA, vol. 21 (2004), pp. 16–21
19. C. Macal, M. North, Introductory tutorial: agent-based modeling and simulation, in *Proceedings of the 2014 Winter Simulation Conference* (IEEE Press, New York, 2014), pp. 6–20
20. D. Morley, K. Myers, The SPARK agent framework, in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, vol. 2, pp. 714–721 (IEEE Computer Society, Washington, 2004)
21. G. An, Q. Mi, J. Dutta-Moscato, Y. Vodovotz, Agent-based models in translational systems biology. Wiley Interdiscip. Rev.: Syst. Biol. Med. **1**(2), 159–171 (2009)
22. P. Kisielow, H. Sia Teh, H. Blüthmann, H. von Boehmer, Positive selection of antigen-specific T cells in thymus by restricting MHC molecules. Nature **335**(6192), 730–733 (1988)
23. S.C. Jameson, K.A. Hogquist, M.J. Bevan, Positive selection of thymocytes. Ann. Rev. Immunol. **13**(1), 93–126 (1995)
24. A. Singer, S. Adoro, J.-H. Park, Lineage fate and intense debate: myths, models and mechanisms of CD4-versus CD8-lineage choice. Nat. Rev. Immunol. **8**(10), 788–801 (2008)
25. M. Sawicka, G.L. Stritesky, J. Reynolds, N. Abourashchi, G. Lythe, C. Molina-París, K.A. Hogquist, From pre-DP, post-DP, SP4, and SP8 thymocyte cell counts to a dynamical model of cortical and medullary selection. Front. Immunol. **5**, 1–14 (2014)
26. T. Hogan, G. Gossel, A.J. Yates, B. Seddon, Temporal fate mapping reveals age-linked heterogeneity in naive T lymphocytes in mice. Proc. Natl. Acad. Sci. **112**(50), E6917–E6926 (2015)
27. T.G. Kurtz, The relationship between stochastic and deterministic models for chemical reactions. J. Chem. Phys. **57**(7), 2976–2978 (1972)
28. D.T. Gillespie, Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81**, 2340–2361 (1977)
29. I. den Braber, T. Mugwagwa, N. Vrisekoop, L. Westera, R. Mögling, A. Bregje de Boer, N. Willems, E.H.R. Schrijver, G. Spierenburg, K. Gaiser, E. Mul, S.A. Otto, A.F.C. Ruiter, M.T. Ackermans, F. Miedema, J.A.M. Borghans, R.J. de Boer, K. Tesselaar, Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. Immunity **36**(2), 288–297 (2012)
30. V.V. Ganusov, R.J. De Boer, Do most lymphocytes in humans really reside in the gut? Trends Immunol. **28**(12), 514–518 (2007)
31. V.V. Ganusov, J.A.M. Borghans, R.J. De Boer, Explicit kinetic heterogeneity: mathematical models for interpretation of deuterium labeling of heterogeneous cell populations. PLoS Comput. Biol. **6**(2), e1000666 (2010)
32. J.J.C. Thome, N. Yudanin, Y. Ohmura, M. Kubota, B. Grinshpun, T. Sathaliyawala, T. Kato, H. Lerner, Y. Shen, D.L. Farber, Spatial map of human T cell compartmentalization and maintenance over decades of life. Cell **159**(4), 814–828 (2014)
33. G. Nigel Gilbert, *Agent-Based Models* (Sage, Beverley Hills, 2008)

Part III
Analysis of Stochastic Dynamical Systems for Modeling Cell Biology

Model Reduction for Stochastic Reaction Systems

Stephen Smith and Ramon Grima

1 Introduction

Master equations constitute the standard description of stochastic reaction dynamics in well-mixed conditions [1]. For linear systems, the moments can be found exactly in closed-form and sometimes even the probability distribution can be obtained [2]. However, most systems are nonlinear, specifically those involving the interaction of two or more entities. For such systems the master equation can be rarely solved and a different solution strategy becomes necessary. A common approach involves performing stochastic simulations using the stochastic simulation algorithm (SSA) [3]. However, the computational expense involved in gathering accurate statistics can be considerable for many systems of interest, particularly those which involve a high number of reaction events per unit time, a feature of systems with one or more abundant species. One means to circumvent this problem involves deriving a reduced master equation for only the non-abundant species and then obtaining the statistics of the number fluctuations using the SSA. Various methods exist which lead to such a reduction [4–7]. Here we choose to focus on what is perhaps the simplest of such methods, one which is easy to derive for all cases of interest and which leads to accurate and fast stochastic simulations. An additional bonus is that in quite a number of cases, the reduced master equation can be solved exactly.

S. Smith • R. Grima (✉)
School of Biological Sciences, University of Edinburgh, Edinburgh, UK
e-mail: ramon.grima@ed.ac.uk

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_7

2 The Method

In this section we introduce the rationale behind the model reduction method, by means of a simple chemical reaction system. For a more technical presentation the reader is referred to the original paper which derives the method [8]. This section is self-contained, assuming no knowledge of rate equations or of master equations, and builds the latter mathematical descriptions and the reduction method from the ground up.

Consider the following chemical reaction system:

$$\emptyset \xrightarrow{k_1} A,\ A + A \xrightarrow{k_2} B,\ B \xrightarrow{k_3} \emptyset. \tag{1}$$

This system consists of two proteins, A and B, and three reactions. The first reaction is the creation of a molecule of A out of nothingness ($\emptyset$). This process is, of course, thermodynamically implausible, but this notation is used when A is created at a more-or-less constant rate by a reaction between chemical species which we are not interested in modelling. The second reaction is between two molecules of A to form a molecule B. We call B a *dimer* and the reaction a *dimerisation*. The third reaction is the destruction of B into nothingness. Again, this is shorthand for the conversion of B into products which we do not want to explicitly model. The quantities $k_1,\ k_2,\ k_3$ are the *rates* of the reactions, i.e. a measure of how frequently they occur.

The most common method of modelling systems like (1) is to treat the concentrations of A and B (number of molecules per unit volume) as continuous functions of time, $[A]$ and $[B]$, respectively. These functions are defined as the solution to a set of differential equations called rate equations (RE):

$$\begin{aligned} \frac{d[A]}{dt} &= k_1 - 2k_2[A]^2, \\ \frac{d[B]}{dt} &= k_2[A]^2 - k_3[B]. \end{aligned} \tag{2}$$

The equation for $\frac{d[A]}{dt}$ states that the rate of change of $[A]$ is equal to the rate of creation of A molecules, k_1, minus the rate of destruction of A molecules, $2k_2[A]^2$. The factor of 2 in the latter term refers to the fact that *two* molecules of A are destroyed whenever the reaction occurs; the factor of $[A]^2$ relates to the fact the reaction happens with a frequency proportional to the number of pairs of A molecules, which scales as $[A]^2$. The equation for $\frac{d[B]}{dt}$ states that the rate of change of $[B]$ is equal to the rate of creation of B molecules, $k_2[A]^2$, minus the rate of destruction of B molecules, $k_3[B]$. The term $k_3[B]$ corresponds to the fact that the third reaction happens with a frequency proportional to the number of B molecules, which scales as $[B]$. For more details, a standard reference book can be consulted [1].

The RE system (2) can be solved for $[A]$ and $[B]$ as functions of time, but for simplicity we will consider the equilibrium (steady-state) case, i.e. when $\frac{d[A]}{dt} = \frac{d[B]}{dt} = 0$. Setting the equations in system (2) equal to zero leads to the following simple expressions:

$$[A] = \sqrt{\frac{k_1}{2k_2}},\ [B] = \frac{k_1}{2k_3}. \tag{3}$$

We therefore know how the concentrations of A and B scale with the various reaction rates. However, our entire line of thinking so far has relied on the assumption that the concentrations of A and B are continuous (even differentiable) functions of time, an assumption which is clearly untrue since the number of molecules of A and B must be integer-valued. Following this line of thinking leads us to consider not the concentrations $[A]$ and $[B]$ but the molecule number n_A and n_B. We furthermore observe that chemical kinetics is not deterministic but rather probabilistic. The reason is that the timing of reaction events is random; for example, the precise time at which two molecules of A will meet is unknown because the process which brings the molecules together, Brownian motion, is a stochastic process. We are therefore concerned with the quantity $P(n_A, n_B; t)$, the probability that our system (1) consists of exactly n_A molecules of A and n_B molecules of B at time t.

Just as with the deterministic RE system (2), the probability $P(n_A, n_B; t)$ is described by a differential equation of the form:

$$\begin{aligned}\frac{d}{dt}P(n_A, n_B; t) &= k_1 V \left(P(n_A - 1, n_B; t) - P(n_A, n_B; t)\right) \\ &+ \frac{k_2}{V}\left((n_A + 2)(n_A + 1)P(n_A + 2, n_B - 1; t) - n_A(n_A - 1)P(n_A, n_B; t)\right) \\ &+ k_3\left((n_B + 1)P(n_A, n_B + 1; t) - n_B P(n_A, n_B; t)\right).\end{aligned} \tag{4}$$

This equation is a master equation, specifically a chemical master equation (CME). It describes the rate of change of the probability of the system having n_A, n_B molecules of A and B, respectively, or in the language of statistical physics, the rate of change of the probability that the system is in the *state* (n_A, n_B).

The first term of Eq. (4) concerns how the system could enter or leave the state (n_A, n_B) due to the first reaction. The system could enter the state (n_A, n_B) due to the production of a molecule of A, if the system was previously in the state $(n_A - 1, n_B)$ [hence the probability $P(n_A - 1, n_B; t)$] or else the system could leave the state (n_A, n_B) if it was already in that state [hence the probability $P(n_A, n_B; t)$]. Note that the first reaction happens with a rate $k_1 V$, for reaction volume V, because the production reaction can occur anywhere throughout the reaction volume.

The second term of Eq. (4) concerns how the system could enter or leave the state (n_A, n_B) due to the second (dimerisation) reaction. The system could enter the state (n_A, n_B) if it was previously in the state $(n_A + 2, n_B - 1)$, or else the system could leave the state (n_A, n_B) if it was already in that state. If the system was in state $(n_A + 2, n_B - 1)$, then the second reaction would occur with rate $\frac{k_2}{V}(n_A + 2)(n_A + 1)$, since the number of distinct pairs of A molecules scales as $(n_A + 2)(n_A + 1)$, and the factor V corresponds to the fact that dimerisation reactions are less likely to occur in large volumes (since molecules are less likely to collide). Similarly, if the system was in state (n_A, n_B), then the second reaction would occur with rate $\frac{k_2}{V} n_A(n_A - 1)$, since the number of distinct pairs of A molecules scales as $n_A(n_A - 1)$.

The third term of Eq. (4) concerns how the system could enter or leave the state (n_A, n_B) due to the third reaction. The system could enter the state (n_A, n_B) if it was previously in state $(n_A, n_B + 1)$, or else the system could leave the state (n_A, n_B) if it was already in that state. If the system was in state $(n_A, n_B + 1)$ the reaction would occur with a rate $k_3(n_B + 1)$, since there are $(n_B + 1)$ molecules of B, but there is no volume dependence for this reaction. Similarly, if the system was in state (n_A, n_B), the third reaction would occur with a rate $k_3 n_B$. For more details, see a standard reference book on stochastic methods [1].

While it can be seen that Eq. (4) corresponds to a physically accurate description of the chemical system (1), it is by no means clear how to solve it, even at steady-state. In fact, Eq. (4) cannot be easily solved analytically. Only a small number of CMEs can be solved, and these generally have special properties such as conservation laws [9], detailed balance [10] or no bimolecular reactions [11]. Instead, the most common stochastic approach to systems like (1) is to simulate them using the stochastic simulation algorithm (SSA) [3]. Performing a large number of independent simulations can generate a histogram which is known to agree with the solution of the CME, within sampling error. However, this technique does not tell us how the average number of molecules of A and B (other moments) depend on the various parameters, analogously to Eqs. (3) for the rate equations. If we want this kind of information we must use methods which analytically approximate $P(n_A, n_B; t)$, and we will describe one such method here.

This method [8] makes the assumption that some of the chemical species have a high concentration. For instance, in system (1), suppose that there are a large number of A molecules. By the definition of concentration, we can approximate the number of A molecules as:

$$n_A \approx V[A], \tag{5}$$

where $[A]$ is the concentration given in Eq. (3). In fact, this approximation becomes more accurate for larger concentrations of $[A]$. As shown in Fig. 1, the stochastic behaviour of n_A (as simulated with the SSA) becomes less distinguishable from the constant solution Eq. (3) as n_A increases. It follows that the stochastic behaviour of n_A becomes less relevant as $[A]$ grows, and it also follows that, if $[A]$ is large,

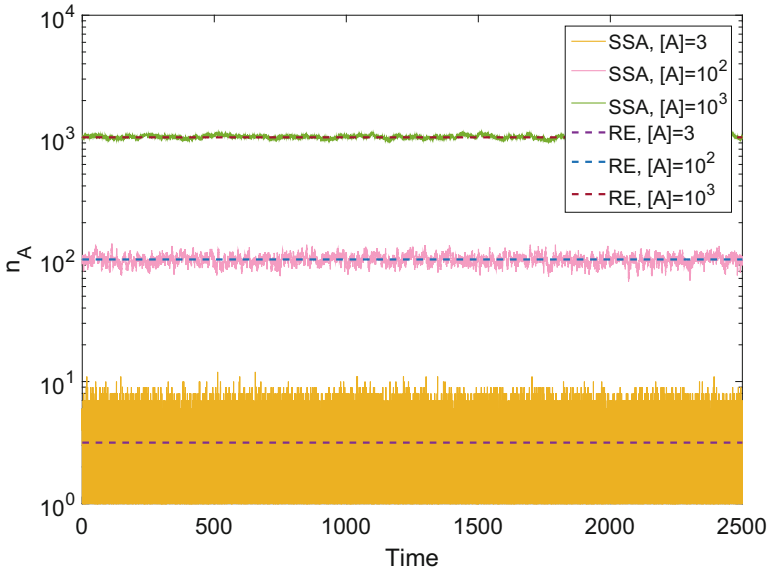


Fig. 1 Stochastic simulations (*yellow*, *pink*, *green*) and deterministic rate-equation solutions (*purple*, *blue*, *red*) for n_A is the system (1). Parameter values are $k_1 = 20$, $k_3 = 1$, $V = 1$ and $k_2 = 1, 10^{-3}, 10^{-5}$ for $[A] = 3, 10^2, 10^3$, respectively. Note that the noise about the mean decreases with increasing mean number of molecules

we can replace every instance of n_A in Eq. (4) with $V[A]$. Also, since n_A is large, $(n_A + i) \approx V[A]$ for small values of i. Hence we have

$$\begin{aligned}\frac{d}{dt}P(V[A], n_B; t) &= k_1 V \left(P(V[A], n_B; t) - P(V[A], n_B; t)\right) \\ &+ \frac{k_2}{V}\left((V[A])^2 P(V[A], n_B - 1; t) - (V[A])^2 P(V[A], n_B; t)\right) \\ &+ k_3 \left((n_B + 1)P(V[A], n_B + 1; t) - n_B P(V[A], n_B; t)\right).\end{aligned} \tag{6}$$

Several simplifications can be made to this equation. The first term, for instance, is now identically zero. Also, notice that since $[A]$ is simply the constant defined in Eq. (3), there is no need to include it as an argument in the probability P, hence we can define a new, simplified probability $\tilde{P}(n_B; t)$, leading to a simplified CME:

$$\begin{aligned}\frac{d}{dt}\tilde{P}(n_B; t) =& \frac{k_1}{2} V \left(\tilde{P}(n_B - 1; t) - \tilde{P}(n_B; t)\right) \\ &+ k_3 \left((n_B + 1)\tilde{P}(n_B + 1; t) - n_B \tilde{P}(n_B; t)\right).\end{aligned} \tag{7}$$

This equation is relatively straightforward, and is known to have the steady-state solution:

$$\tilde{P}(n_B) = \frac{e^{-\frac{k_1 V}{2k_3}} \left(\frac{k_1 V}{2k_3}\right)^{n_B}}{n_B!}. \tag{8}$$

In other words, the number of B molecules approximately satisfies a Poisson distribution with mean $k_1 V/2k_3$.

This approximation can also be made in a simpler (albeit less rigorous) manner immediately from the reaction system (1). As a rule, wherever we see A to the right of an arrow, we simply delete it (replace it with $\emptyset$). Wherever we see A to the left of an arrow, we absorb it as $[A]$ into the reaction rate. This replaces system (1) with the system:

$$\emptyset \xrightarrow{k_1} \emptyset, \ \emptyset \xrightarrow{k_2[A]^2} B, \ B \xrightarrow{k_3} \emptyset. \tag{9}$$

The first reaction here is clearly pointless, and $[A]$ can be replaced by its value given in Eq. (3) to give:

$$\emptyset \xrightarrow{\frac{k_1}{2}} B, \ B \xrightarrow{k_3} \emptyset. \tag{10}$$

It can be shown that the CME for system (10) is identically Eq. (7). We therefore have a very quick way to reduce complex chemical reaction systems to simpler subsystems which we know how to solve.

Yet, although this technique has been derived a priori, a question remains as to the accuracy of the simplified system. How large does $[A]$ have to be before the simplified system is a good approximation to the true system? In fact, for most cases, the simplified system is reasonable even when $[A]$ is not large. For example, in Fig. 2 we show how the Poisson distribution in Eq. (8) agrees very well with the true solution of the CME when $[A]$ is large, but also shows reasonable agreement when $[A]$ is much smaller than $[B]$. In general, however, the error scales as the inverse ratio of the two concentrations, so that the larger $[A]$ is compared to $[B]$, the better the approximation.

The example shown above is illustrative but of limited biological relevance or interest. Next we will therefore demonstrate the power of the analytical approximation method by applying it to a variety of biological systems.

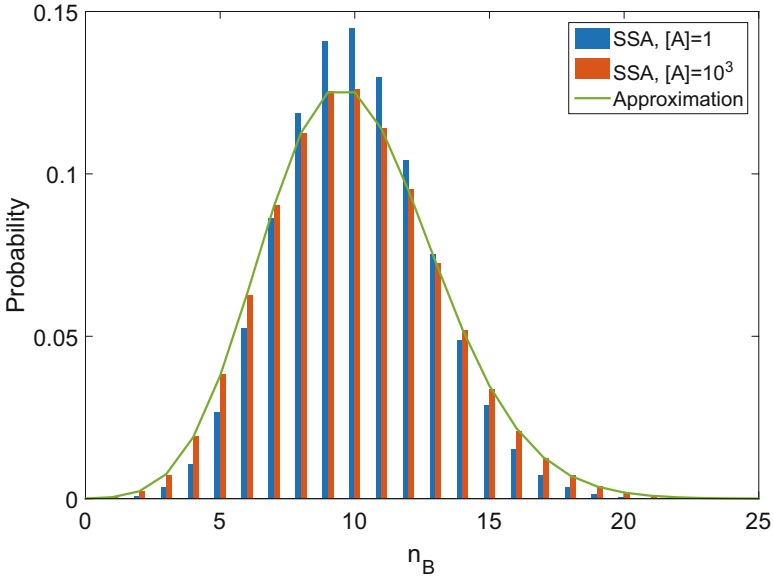


Fig. 2 Probability distributions of n_B obtained from the SSA (histograms) and the analytical approximation assuming abundance of A as given by Eq. (8) (*solid line*). Parameter values are as in Fig. 1 and $k_2 = 10, 10^{-5}$ for $[A] = 1, 10^3$, respectively. Note that the approximation is particularly good when $[A] \gg [B] = 10$

3 Application to Various Biological Systems

3.1 Michaelis–Menten Reaction

Consider the set of chemical reactions:

$$\emptyset \xrightarrow{k_1} S,\ E + S \xrightarrow{k_2} C,\ C \xrightarrow{k_3} E + S,\ C \xrightarrow{k_4} E + P. \tag{11}$$

The system describes how a protein S is created (with rate k_1) and reacts with an enzyme E to form a complex C (with rate k_2). C can subsequently unbind, either back to $E + S$ (with rate k_3) or else to E and new protein P (with rate k_4). This system has been studied extensively with and without protein production, using rate equations and master equations [12, 13]. The first reaction could, for example, model the effective translation of a protein in the cytoplasm while the rest of the reactions model the enzyme-aided catalysis of the protein into another type of protein.

The REs for this system are given by:

$$\begin{aligned}\frac{d[S]}{dt} &= k_1 - k_2[E][S] + k_3[C],\\ \frac{d[E]}{dt} &= -k_2[E][S] + (k_3 + k_4)[C],\\ \frac{d[C]}{dt} &= k_2[E][S] - (k_3 + k_4)[C].\end{aligned} \tag{12}$$

In steady-state, the solutions of this system are

$$[S] = \frac{k_1(k_3 + k_4)}{k_2(E_T k_4 - k_1)},\ [E] = E_T - \frac{k_1}{k_4},\ [C] = \frac{k_1}{k_4}, \tag{13}$$

where $E_T = [E] + [C]$ is the total enzyme concentration which remains constant over time. In order to apply the approximation method, we make the assumption that $[S]$ is large compared to $[E]$ and $[C]$. Then, we can reduce the reaction network to give:

$$E \xrightarrow{k_2[S]} C,\ C \xrightarrow{k_3+k_4} E. \tag{14}$$

The steady-state solution of the chemical master equation for this effective system of reactions is easily found to be a Binomial $\left(E_T V, \frac{k_3+k_4}{k_3+k_4+k_2[S]}\right)$ distribution for n_E. In Fig. 3 we compare this Binomial distribution with SSA histograms for three different values of $[S]$. As $[S]$ increases, it is clear that the approximation improves, though the Binomial distribution seems to be a reasonable estimate of the distribution even when $[S]$ is comparable to E.

3.2 Genetic Network with Feedback

Next we study a simple model of a genetic network with negative feedback:

$$\begin{aligned}&D_{\mathrm{ON}} \xrightarrow{k_1} D_{\mathrm{ON}} + M,\ M \xrightarrow{k_2} M + P_1,\ P_1 \xrightarrow{k_3} P_2,\ P_2 \xrightarrow{k_4} P_3,\ P_3 \xrightarrow{k_5} \emptyset,\\ &D_{\mathrm{ON}} + P_3 \xrightarrow{k_6} D_{\mathrm{OFF}},\ D_{\mathrm{OFF}} \xrightarrow{k_7} D_{\mathrm{ON}} + P_3,\ M \xrightarrow{k_8} \emptyset.\end{aligned} \tag{15}$$

The system describes how a molecule of mRNA M is produced by an active gene D_{ON} with rate k_1 (transcription), which in turn creates a protein P_1 with rate k_2 via

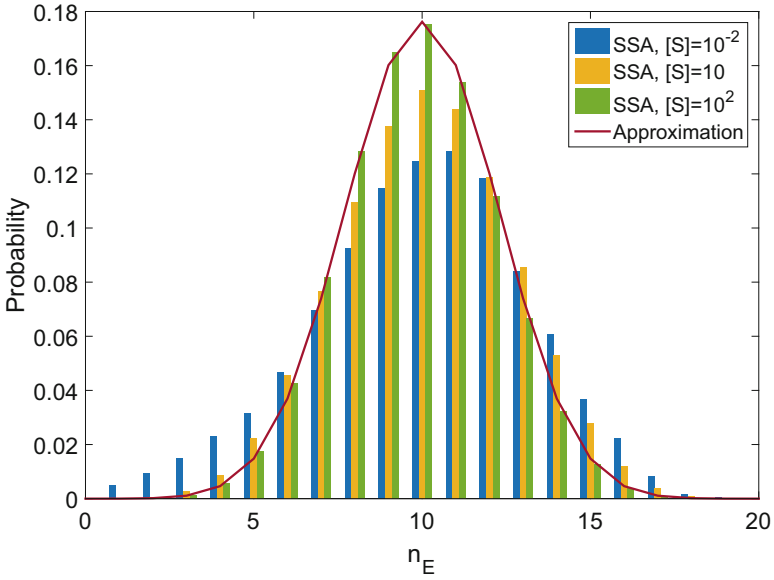


Fig. 3 Probability distribution of n_E is the Michaelis–Menten reaction system (11) obtained from the SSA of the full system (histograms) and the analytical approximation corresponding to the reduced system (14) assuming abundance of S compared to enzyme (*red solid line*). Parameter values are $k_1 = 1,\ k_3 = 9.9,\ k_4 = 0.1,\ E_T = 20,\ V = 1$. k_2 is varied to give different values of $[S]$. Note that the approximation is excellent when the concentration of S is much greater than that of enzyme

the process of translation. P_1 isomerises to P_2 with rate k_3 which isomerises to P_3 with rate k_4. P_3 can decay with rate k_5 or it can bind to the active gene D_{ON} with rate k_6 to convert it to the inactive gene D_{OFF}, and the deactivation can be reversed with rate k_7. Finally, the mRNA can decay with rate k_8. The system possesses negative feedback since the protein produced by the gene in the on state can turn the gene off at high enough concentrations; it has previously been studied as a simple model of a circadian oscillator [14, 15].

The REs for this system are given by:

$$\frac{d[D_{\text{ON}}]}{dt} = -k_6[D_{\text{ON}}][P_3] + k_7[D_{\text{OFF}}],$$

$$\frac{d[D_{\text{OFF}}]}{dt} = k_6[D_{\text{ON}}][P_3] - k_7[D_{\text{OFF}}],$$

$$\frac{d[M]}{dt} = k_1[D_{\text{ON}}] - k_8[M],$$

$$\frac{d[P_1]}{dt} = k_2[M] - k_3[P_1],$$
$$\frac{d[P_2]}{dt} = k_3[P_1] - k_4[P_2],$$
$$\frac{d[P_3]}{dt} = k_4[P_2] - k_5[P_3] - k_6[D_{\text{ON}}][P_3] + k_7[D_{\text{OFF}}]. \tag{16}$$

This system of equations can be solved in steady-state, but the expressions are cumbersome and we will not state them here except for $[P_3]$:

$$[P_3] = \frac{-k_7 + \sqrt{k_7^2 + \dfrac{4k_6k_1k_2k_7N}{k_5k_8}}}{2k_6}, \tag{17}$$

where $N = [D_{\text{ON}}] + [D_{\text{OFF}}]$ is the total gene concentration which remains constant over time. In order to apply the approximation method, we make the reasonable assumption that the protein concentrations $[P_1]$, $[P_2]$, $[P_3]$ are large compared to $[D_{\text{ON}}]$, $[D_{\text{OFF}}]$ and $[M]$. For example, it has been shown that the mean number of proteins per *E. coli* cell is roughly a thousand times that of the mean number of mRNA molecules per cell while the gene copy number is often one [16]. Then, we can reduce the reaction network to give:

$$D_{\text{ON}} \xrightarrow{k_1} D_{\text{ON}} + M, D_{\text{ON}} \xrightarrow{k_6[P_3]} D_{\text{OFF}},\ D_{\text{OFF}} \xrightarrow{k_7} D_{\text{ON}},\ M \xrightarrow{k_8} \emptyset. \tag{18}$$

The chemical master equation for this system can be solved exactly in steady-state using the generating function method [9] . The solution is complex so we will not state it here, but in Fig. 4 we compare the analytical approximation with the SSA distribution, and observe that they agree well. In the inset we show trajectories of the SSA for M (red) and P_1, P_2, P_3 (yellow, blue, purple), which highlight the bimodal nature of this system. The parameter set chosen here highlights two remarkable properties of the approximation method. First, the approximate distribution captures the bimodality of the true distribution. This is surprising because approximation methods are rarely able to deal with bimodality. Second, the distribution is accurate even though for a significant portion of time the system is in the lower state where $[P_1] = [P_2] = [P_3] = 0$, which can hardly be described as a high concentration. It can be shown that the accuracy of the approximation depends only on whether the RE solution is large, irrespective of the stochastic behaviour.

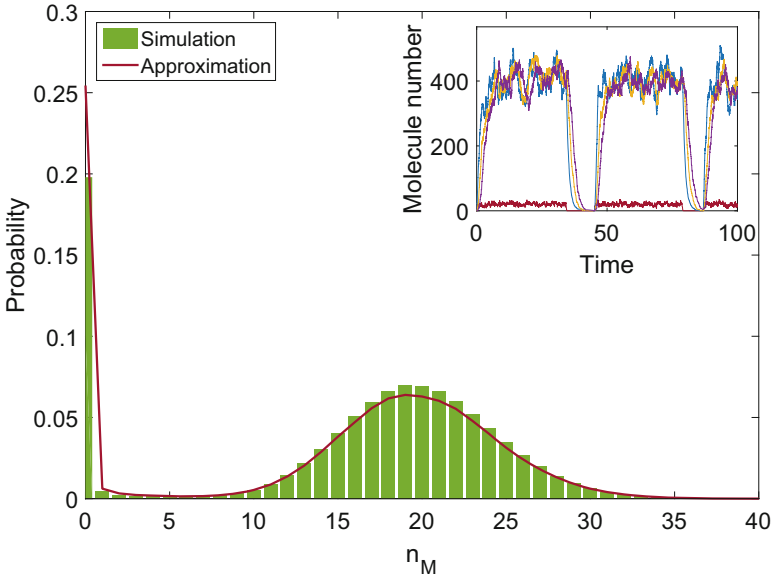


Fig. 4 Probability distribution of the genetic feedback system (15), as obtained from the SSA of the full system (histogram) and the approximation corresponding to the reduced system (18) assuming protein abundance (*red solid line*). Inset: time courses of M (*red*) and P_1, P_2, P_3 (*blue*, *yellow*, *purple*) as obtained from SSA, showing bimodality in both mRNA and protein values and abundance of protein compared to mRNA for most of the time. Parameter values are $k_1 = 100$, $k_2 = 5$, $k_3 = 20$, $k_4 = 1$, $k_5 = 1$, $k_6 = 1$, $k_7 = 10^{-4}$, $k_8 = 0.1$, $N = 1$, $V = 1$

3.3 *Biochemical Switch*

Next we consider a simple model of a biochemical switch [17], which could be used to construct synthetic logic gates:

$$C \xrightarrow{k_1} C + X_1,\ X_1 \xrightarrow{k_2} \emptyset,\ X_1 + B \underset{k_4}{\overset{k_3}{\rightleftharpoons}} A,\ X_2 + A \underset{k_6}{\overset{k_5}{\rightleftharpoons}} B,\ \emptyset \underset{k_8}{\overset{k_7}{\rightleftharpoons}} X_2. \tag{19}$$

The system describes how three enzymes A, B, C catalyse the production and degradation of two proteins X_1 and X_2. The engineering application of this system is that the output, A, is an amplification of the input, C, since a small change in $[C]$

corresponds to a large change in $[A]$, thus functioning as a switch. The REs for this system are given by:

$$\begin{aligned}
\frac{d[A]}{dt} &= k_3[X_1][B] - k_4[A] - k_5[X_2][A] + k_6[B], \\
\frac{d[B]}{dt} &= -k_3[X_1][B] + k_4[A] + k_5[X_2][A] - k_6[B], \\
\frac{d[X_1]}{dt} &= k_1[C] - k_2[X_1] - k_3[X_1][B] + k_4[A], \\
\frac{d[X_2]}{dt} &= -k_5[X_2][A] + k_6[B] + k_7 - k_8[X_2].
\end{aligned} \tag{20}$$

This system can be solved in steady-state, but the expressions are cumbersome and we will not state them here. In order to apply the approximation method, we make the reasonable assumption that the protein concentrations $[X_1]$, $[X_2]$ are large compared to the enzyme concentrations $[A]$ and $[B]$ (similar to the assumption made for the Michaelis–Menten system). Then, we can reduce the reaction network to the effective one:

$$A \underset{k_3[X_1]+k_6}{\overset{k_4+k_5[X_2]}{\rightleftharpoons}} B. \tag{21}$$

Analogously to system (14), the chemical master equation for this system has a steady-state solution given by a Binomial distribution, specifically a Binomial $\left(E_T V, \frac{k_4+k_5[X_2]}{k_4+k_5[X_2]+k_3[X_1]+k_6}\right)$ distribution for n_A, where $E_T = [A] + [B]$ is the total concentrations of enzyme, which remains constant over time.

In Fig. 5 we plot the distribution of n_A for a variety of values of $[C]$, for both the analytical approximation corresponding to the effective system (21) and stochastic simulations using the SSA of the master equation of the full system (19). The switch-like behaviour is clear to see, as the mean of n_A moves swiftly from about 10 to near 100 molecules (almost a ninefold change) as $[C]V$ (the input number of C molecules) is changed from 500 to 1500 molecules (a twofold change). The approximation method here provides a very quick means of modelling a bioengineering component, which would otherwise be time-consuming to simulate.

3.4 Predator–Prey System

Lastly we consider an example from ecology, a Lotka–Volterra-like predator–prey system [18] given by the reactions:

$$R \xrightarrow{k_1} R + R,\ R + F \xrightarrow{k_2} F + F,\ F \xrightarrow{k_3} \emptyset,\ \emptyset \xrightarrow{k_4} R,\ \emptyset \xrightarrow{k_5} F. \tag{22}$$

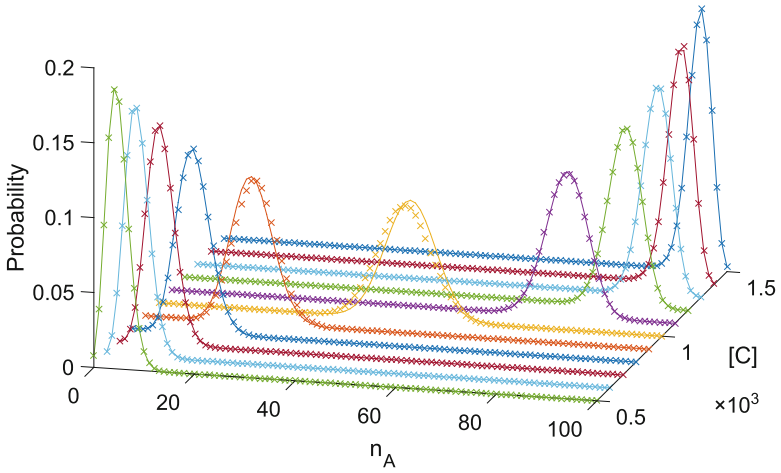


Fig. 5 Probability distributions of n_A in the biochemical switch (19) for different values of $[C]$, obtained from the SSA of the full system (*crosses*) and the analytical approximation corresponding to the reduced system (21) under the assumption of protein abundance (*solid lines*). Parameter values are $k_1 = 0.05$, $k_2 = 0.001$, $k_3 = 0.001$, $k_4 = 1$, $k_5 = 0.001$, $k_6 = 1$, $k_7 = 0.001$, $k_8 = 50$, $V = 1$. Note that a twofold change in the input [C] leads to an almost ninefold change in the mean of the output A

The system describes (in a simplistic way) how a population of foxes F predates on a population of rabbits R. A single rabbit can reproduce with rate k_1 and a fox can eat a rabbit, giving it enough energy to reproduce with rate k_2. Foxes can die with rate k_3, and rabbits and foxes can both immigrate into the environment with rates k_4 and k_5, respectively. The REs for this system are

$$\begin{aligned} \frac{d[R]}{dt} &= k_1[R] - k_2[R][F] + k_4, \\ \frac{d[F]}{dt} &= k_2[R][F] - k_3[F] + k_5. \end{aligned} \tag{23}$$

This system of equations can exhibit oscillatory behaviour for some parameter values and stable behaviour for others.

$$\begin{aligned} [R] &= \frac{2k_3k_4}{k_2k_4 - k_1k_3 + k_2k_5 + \sqrt{(k_2k_4 + k_1k_3 + k_2k_5)^2 - 4k_1k_2k_3k_5}}, \\ [F] &= \frac{k_2k_4 + k_1k_3 + k_2k_5 + \sqrt{(k_2k_4 + k_1k_3 + k_2k_5)^2 - 4k_1k_2k_3k_5}}{2k_2k_3}. \end{aligned} \tag{24}$$

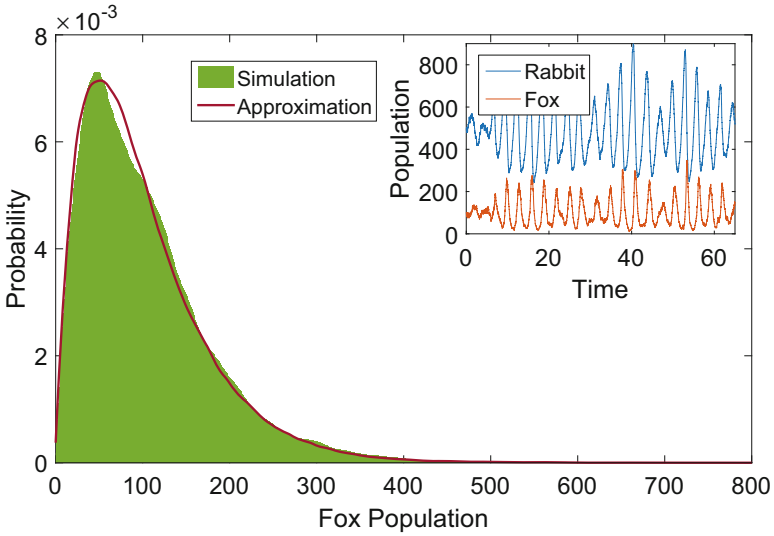


Fig. 6 Probability distribution of the number of foxes, n_F, in the predator–prey model (22) as obtained from the SSA of the full system (histogram) and the analytical approximation corresponding to the reduced system (25) which assumes an abundance of rabbits (*solid line*). Inset: time course data of rabbit (*blue*) and fox (*red*) populations from the SSA of the full system. Parameter values are $k_1 = 1,\ k_2 = 0.01,\ k_3 = 5,\ k_4 = 5,\ k_5 = 10,\ V = 1$

In order to apply the approximation method, we make the reasonable assumption that the concentration of rabbits $[R]$ is large compared to the concentration of foxes $[F]$. Then, we can reduce the reaction network to give:

$$F \xrightarrow{k_2[R]} F + F,\ F \xrightarrow{k_3} \emptyset,\ \emptyset \xrightarrow{k_5} F. \tag{25}$$

The master equation for this effective system of reactions can be solved analytically in steady-state [19] and its solution is a Negative Binomial $\left(\frac{k_5}{k_2[R]}, \frac{k_2[R]}{k_3}\right)$. In Fig. 6 we compare this analytical approximation with the distribution obtained using the SSA of the full system (22), and note that they agree well. In the inset we show a time course of the SSA of the full system for rabbits (blue) and foxes (red). Note that the oscillations are here induced by noise and are not predicted by the deterministic rate equations. It can be shown that the reduced system does not possess noise-induced oscillations; this is because for a one variable system the power spectrum of noise fluctuations cannot exhibit a peak at a non-zero frequency [20]. Hence while the approximation method leads to an excellent approximation for the probability distribution of the full system it cannot always capture other relevant stochastic properties.

4 Conclusion

In this chapter, we have introduced a simple technique of approximating the master equation. This involves reducing the model to one with a smaller number of species interacting via a set of effective reactions. If the number of species in the reduced model is small, then there is a good chance of finding an analytical solution in steady-state, as we have seen for many examples. Even if such an explicit closed-form solution is not possible, stochastic simulation using the SSA of the reduced master equation yields an accurate solution in a time which is typically much shorter than that possible with the SSA of the full master equation.

References

1. N. van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd edn. (Elsevier, Amsterdam, 2007)
2. C. Gadgil, C. Lee, H. Othmer, A stochastic analysis of first-order reaction networks. Bull. Math. Biol. **67**, 901–946 (2005)
3. D.T. Gillespie, Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81**, 2340–2361 (1977)
4. A. Hellander, P. Lötstedt, Hybrid method for the chemical master equation. J. Comput. Phys. **227**, 100–122 (2007)
5. S. Peleš, B. Munsky, M. Khammash, Reduction and solution of the chemical master equation using time scale separation and finite state projection. J. Chem. Phys. **125**, 204104 (2006)
6. P. Thomas, A.V. Straube, R. Grima, The slow-scale linear noise approximation: an accurate, reduced stochastic description of biochemical networks under timescale separation conditions. BMC Syst. Biol. **6**, 39 (2012)
7. T. Jahnke, On reduced models for the chemical master equation. SIAM Multiscale Model. Simul. **9**, 1646–1676 (2011)
8. S. Smith, C. Cianci, R. Grima, Model reduction for stochastic chemical systems with abundant species. J. Chem. Phys. **143**, 214105 (2015)
9. R. Grima, D. Schmidt, T. Newman, Steady-state fluctuations of a genetic feedback loop: an exact solution. J. Chem. Phys. **137**, 035104 (2012)
10. C. Cianci, S. Smith, R. Grima, Molecular finite-size effects in stochastic models of equilibrium chemical systems. J. Chem. Phys. **144**, 084101 (2016)
11. V. Shahrezaei, P.S. Swain, Analytical distributions for stochastic gene expression. Proc. Natl. Acad. Sci. **105**, 17256–17261 (2008)
12. S. Schnell, C. Mendoza, Closed form solution for time-dependent enzyme kinetics. J. Theor. Biol. **187**, 207–212 (1997)
13. R. Grima, Noise-induced breakdown of the Michaelis-Menten equation in steady-state conditions. Phys. Rev. Lett. **102**, 218103 (2009)
14. D. Gonze, J. Halloy, A. Goldbeter, Robustness of circadian rhythms with respect to molecular noise. Proc. Natl. Acad. Sci. **99**, 673–678 (2002)
15. P. Thomas, A.V. Straube, J. Timmer, C. Fleck, R. Grima, Signatures of nonlinearity in single cell noise-induced oscillations. J. Theor. Biol. **335**, 222–234 (2013)
16. Y. Taniguchi, P. Choi, G. Li, H. Chen, M. Babu, J. Hearn, A. Emili, X.S. Xie, Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science **329**, 533–538 (2010)

17. J. Ross, I. Schreiber, V. MO, *Determination of Complex Reaction Mechanisms: Analysis of Chemical, Biological, and Genetic Networks* (Oxford University Press, Oxford, 2005)
18. J. Murray, *Mathematical Biology I: An Introduction* (Springer, New York, 2002)
19. S. Smith, V. Shahrezaei, General transient solution of the one-step master equation in one dimension. Phys. Rev. E **91**, 062119 (2015)
20. C. Gardiner, *Handbook of Stochastic Methods* (Springer, Berlin, 1985)

ZI-Closure Scheme: A Method to Solve and Study Stochastic Reaction Networks

M. Vlysidis, P.H. Constantino, and Y.N. Kaznessis

1 Introduction

We use an example to present in exhaustive detail the algorithmic steps of the zero-information (ZI) closure scheme [1]. ZI-closure is a method for solving the chemical master equation (CME) of stochastic chemical reaction networks.

The objective of this chapter is twofold: first to present the algorithm with sufficient didactic value for a non-expert, yet patient and self-motivated reader to confidently reconstruct the algorithm with any available programming tools, including the ubiquitous paper and pencil. Second, this chapter lays bare the theoretical and numerical underpinnings of ZI-closure for the expert audience to confidently criticize and propose new, improved implementation solutions.

An avalanche of text, some of it wonderful and inspiring, has been written in the last two decades on the importance of stochasticity in chemical reaction networks [2–12]. We assume that the reader is familiar with this literature. The historical difficulties in solving the CME are well documented in the literature. The CME is not a single equation but a set of infinite number of coupled equations. As a result the CME is analytically unsolvable. One of the most popular methods to numerically solve the CME is Gillespie's stochastic simulation algorithm [9]. However, the method is a kinetic Monte Carlo algorithm and hence computationally expensive. ZI-closure scheme is able to solve the CME as accurate as Gillespie's algorithm with significantly less computational time needed.

The starting point of ZI-closure scheme is to rewrite the CME in terms of moments describing the master probability distribution. Instead of a master equation that governs the probability distribution in time, one can then write a set of

M. Vlysidis • P.H. Constantino • Y.N. Kaznessis (✉)
Department of Chemical Engineering and Materials Science, University of Minnesota, 421 Washington Ave. SE, Minneapolis, MN 55455, USA
e-mail: yiannis@umn.edu

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_8

differential equations evolving the expected values of the distribution [13]. Efficient algorithms have been developed to generate these moment equations for arbitrary networks [14]. Yet, because the dynamics of lower-order moments depends on the higher ones for nonlinear reaction networks, the system of ODEs needs to be closed or somehow truncated in order to be solved [15].

At this point, ZI-closure relies on the maximum entropy principle, which states that if some features of a probability distribution are known a priori, then in order to estimate the full distribution we can look for the one with maximum uncertainty that also satisfies the given knowledge (unbiased estimate) [12, 16]. Therefore, using the definition of information entropy by Shannon [17], the system of moment equations is closed by obtaining the higher moments from the maximum information entropy distribution.

Herein we unpack in full mathematical detail the ZI-closure scheme for the system of moment equations. We apply it to a specific reaction network, the Schlögl model, as an illustrative study case. The Schlögl model is a single-component theoretical system that can exhibit bistability. The simplicity of the state space (single-component) alongside with the complexity of the probability distribution (bistability) makes the model a compelling studying example for the ZI-closure scheme.

We start with the analytical derivation of the chemical master equation of the Schlögl model. We then present the derivation of the associated moment equations. Next, we move on to explicitly discuss the theoretical framework supporting the zero-information closure scheme. After presenting a numerical procedure for the implementation of the method, we show some results comparing its accuracy to kinetic Monte Carlo simulations.

2 Theoretical Background

2.1 Derivation of Chemical Master Equation for the Schlögl Model

This section presents the derivation of the chemical master equation for the Schlögl model. The CME models the probability $P(X;t)$ that the system is in state X at time t. Here X represents the number of molecules of the sole reactant X.

The Schlögl model consists of the following reactions [18]:

$$3X \xrightarrow{k_1} 2X, \tag{1}$$

$$2X \xrightarrow{k_2} 3X, \tag{2}$$

$$X \xrightarrow{k_3} \emptyset, \tag{3}$$

$$\emptyset \xrightarrow{k_4} X. \tag{4}$$

The probability of reaction (1) occurring during an infinitesimal time interval $(t, t + dt)$ is $P_1(X) = k_1\, X(X-1)(X-2)\, dt$. For reaction (2) this probability is $P_2(X) = k_2\, X(X-1)\, dt$, for reaction (3) we obtain $P_3(X) = k_3\, X\, dt$ and for reaction (4) the probability is $P_4(X) = k_4\, dt$ [19], with P_i, $i = 1, 2, 3, 4$, being the probability of reaction i firing in isolation. In order for the process to be considered Markovian, at most one reaction is allowed per time interval dt [19].

In the Schlögl model, the only ways for the system to be at X number of molecules at time $t + dt$ are:

(a) The system was at X number of molecules at time t and no reaction occurred within $(t, t + dt)$,
(b) the system had $X - 1$ number of molecules at time t and either reaction (2) or reaction (4) occurred within $(t, t + dt)$,
(c) the system had $X + 1$ number of molecules at time t and either reaction (1) or reaction (3) occurred within $(t, t + dt)$.

The probability of event (a) is:

$$\begin{aligned} P(a) &= P(\text{no reaction occurs}) \cdot P(\text{system has } X \text{ molecules at time } t) \\ &= [1 - P(\text{all reactions occur})]\, P(X; t). \end{aligned}$$

Since all reactions are considered independent, this results in:

$$\begin{aligned} P(a) &= [1 - P_1(X) - P_2(X) - P_3(X) - P_4(X)]\, P(X; t) \\ &= \left[1 - k_1\, X(X-1)(X-2)\, dt - k_2\, X(X-1)\, dt - k_3\, X\, dt - k_4\, dt\right] P(X; t). \end{aligned} \tag{5}$$

Similarly, the probability of event (b) is:

$$\begin{aligned} P(b) &= [P_2(X-1) + P_4(X-1)]\, P(X-1; t) \\ &= \left[k_2\, (X-1)(X-2)\, dt + k_4\, dt\right] P(X-1; t). \end{aligned} \tag{6}$$

And for event (c) we have:

$$\begin{aligned} P(c) &= [P_1(X+1) + P_3(X+1)]\, P(X+1; t) \\ &= \left[k_1\, (X+1)X(X-1)\, dt + k_3\, (X+1)\, dt\right] P(X+1; t). \end{aligned} \tag{7}$$

Thus, the probability that the system is at state X at time $t + dt$ is:

$$
\begin{aligned}
P(X;t+dt) &= P(a) \;+\; P(b) \;+\; P(c) \\
&= +\Big[1 - k_1X(X-1)(X-2)dt - k_2X(X-1)dt - k_3Xdt - k_4\,dt\Big]P(X;t) \\
&\quad + \Big[k_2\,(X-1)(X-2)\,dt + k_4\,dt\Big]\,P(X-1;t) \\
&\quad + \Big[k_1\,(X+1)X(X-1)\,dt + k_3\,(X+1)\,dt\Big]\,P(X+1;t).
\end{aligned}
$$

Rearranging the last expression:

$$
\begin{aligned}
P(X;t+dt) - P(X;t) \;&=\; \Big\{ -\Big[k_1\,X(X-1)(X-2) \;+k_2\,X(X-1) \;+k_3\,X \;+k_4\Big] \\
&\qquad \times P(X;t) + \Big[k_2\,(X-1)(X-2) \;+k_4\Big]\,P(X-1;t) \\
&\qquad + \Big[k_1\,(X+1)X(X-1) \;+k_3\,(X+1)\Big]\,P(X+1;t)\Big\}\,dt \\
&\Longleftrightarrow \frac{P(X;t+dt) - P(X;t)}{dt} \\
&= -\Big[k_1\,X(X-1)(X-2) + k_2\,X(X-1) \;+k_3\,X \;+k_4\Big]\,P(X;t) \\
&\qquad + \Big[k_2\,(X-1)(X-2) \;+k_4\Big]\,P(X-1;t) \\
&\qquad + \Big[k_1\,(X+1)X(X-1) \;+k_3\,(X+1)\Big]\,P(X+1;t).
\end{aligned}
$$

And then by taking the limit $dt \longrightarrow 0$ we have:

$$
\begin{aligned}
&\lim_{dt\to 0} \frac{P(X;t+dt) - P(X;t)}{dt} \\
&= \frac{\partial P(X;t)}{\partial t} = -\Big[k_1\,X(X-1)(X-2) \;+k_2\,X(X-1) \;+k_3\,X \;+k_4\Big]\,P(X;t) \\
&\quad + \Big[k_2\,(X-1)(X-2) \;+k_4\Big]\,P(X-1;t) \\
&\quad + \Big[k_1\,(X+1)X(X-1) \;+k_3\,(X+1)\Big]\,P(X+1;t). \qquad (8)
\end{aligned}
$$

This is the CME for the Schlögl reaction model.

2.2 Derivation of Moment Equations from Chemical Master Equation

A proposed technique to solve the CME relies on calculating the probability moments. The approach is based on the idea that any probability distribution can be completely described by its moments. Moments are expected values of functions of an independent random variable, here the number of molecules, e.g., the mean, the variance, the skewness, and the kurtosis of a probability distribution.

There exist numerous different types of moments, all valid for describing a probability distribution, e.g., the central, factorial, and polynomial probability distribution moments [7, 14, 15, 20].

Herein we use factorial moments given by the expression:

$$\{X^m\} = \sum_{x=0}^{\infty} \frac{x!}{(x-m)!} P(x,t), \tag{9}$$

where x represents the possible values of X and m is the integer order of the moment.

A main reason for using factorial moments is that they are simple derivatives of a Z-transform of the probability distribution, which has the following form:

$$G(S,t) = \sum_{x=0}^{\infty} S^x P(x,t), \tag{10}$$

where S is the continuous transformation of X. Z-transform is also known as the probability-generating function.

Differentiating Eq. (10) with respect to S, at $S = 1$, yields:

$$\left.\frac{\partial G(S,t)}{\partial S}\right|_{S=1} = \{X\}, \qquad \left.\frac{\partial^2 G(S,t)}{\partial S^2}\right|_{S=1} = \{X^2\},$$
$$\left.\frac{\partial^3 G(S,t)}{\partial S^3}\right|_{S=1} = \{X^3\}, \qquad \left.\frac{\partial^4 G(S,t)}{\partial S^4}\right|_{S=1} = \{X^4\}. \tag{11}$$

Similar equations hold for the rest of the moments (fifth-order, sixth-order, etc.). Note that it is also true that $G(S,t)|_{S=1} = 1$. Thus, it is easy to generate factorial moments through the Z-transform of the probability distribution.

Differentiating the function $G(S,t)$ with respect to time, we get:

$$\frac{\partial G(S,t)}{\partial t} = \sum_{x=0}^{\infty} S^x \frac{\partial P(x,t)}{\partial t}. \tag{12}$$

Applying this to the CME for the Schlögl model (8), we obtain the Z-transformed chemical master equation (Z-CME)

$$\frac{\partial G(S,t)}{\partial t} = k_1\left(S^2 - S^3\right)\frac{\partial^3 G}{\partial S^3} + k_2\left(S^3 - S^2\right)\frac{\partial^2 G}{\partial s^2} + k_3(1-S)\frac{\partial G}{\partial S} + k_4(S-1)G. \tag{13}$$

Through Z-CME one can generate moment equations. Taking the first derivative of Eq. (13) with respect to S yields

$$\begin{aligned}\frac{\partial}{\partial t}\left(\frac{\partial G}{\partial S}\right) = \frac{\partial}{\partial S}\left(\frac{\partial G}{\partial t}\right) &= k_1\left(2S - 3S^2\right)\frac{\partial^3 G}{\partial S^3} + k_1\left(S^2 - S^3\right)\frac{\partial^4 G}{\partial S^4} \\ &+ k_2\left(3S^2 - 2S\right)\frac{\partial^2 G}{\partial s^2} + k_2\left(S^3 - S^2\right)\frac{\partial^3 G}{\partial s^3} + k_3(-1)\frac{\partial G}{\partial S} \\ &+ k_3(1-S)\frac{\partial^2 G}{\partial S^2} + k_4(1)G + k_4(S-1)\frac{\partial G}{\partial S}.\end{aligned} \tag{14}$$

Setting $S = 1$ we get

$$\left.\frac{\partial}{\partial t}\left(\frac{\partial G}{\partial S}\right)\right|_{S=1} = -k_1\left.\frac{\partial^3 G}{\partial S^3}\right|_{S=1} + k_2\left.\frac{\partial^2 G}{\partial s^2}\right|_{S=1} - k_3\left.\frac{\partial G}{\partial S}\right|_{S=1} + k_4\left.G\right|_{S=1}. \tag{15}$$

Or from Eq. (11)

$$\frac{\partial\{X\}}{\partial t} = -k_1\{X^3\} + k_2\{X^2\} - k_3\{X\} + k_4. \tag{16}$$

Equation (16) is the equation for the first-order factorial moment of $P(X,t)$.

In order to generate the moment equation for the second moment one needs to take the second derivative of Eq. (13) with respect to S [or the derivative of Eq. (14)]

$$\begin{aligned}\frac{\partial}{\partial t}\left(\frac{\partial^2 G}{\partial S^2}\right) = \frac{\partial^2}{\partial S^2}\left(\frac{\partial G}{\partial t}\right) &= k_1(2-6S)\frac{\partial^3 G}{\partial S^3} + k_1\left(2S - 3S^2\right)\frac{\partial^4 G}{\partial S^4} \\ &+ k_1\left(2S - 3S^2\right)\frac{\partial^4 G}{\partial S^4} + k_1\left(S^2 - S^3\right)\frac{\partial^5 G}{\partial S^5} + k_2(6S-2)\frac{\partial^2 G}{\partial s^2} \\ &+ k_2\left(3S^2 - 2S\right)\frac{\partial^3 G}{\partial s^3} + k_2\left(3S^2 - 2S\right)\frac{\partial^3 G}{\partial s^3} + k_2\left(S^3 - S^2\right)\frac{\partial^4 G}{\partial s^4} - k_3\frac{\partial^2 G}{\partial S^2} \\ &+ k_3(-1)\frac{\partial^2 G}{\partial S^2} + k_3(1-S)\frac{\partial^3 G}{\partial S^3} + k_4\frac{\partial G}{\partial S} + k_4(1)\frac{\partial G}{\partial S} + k_4(S-1)\frac{\partial^2 G}{\partial S^2}\end{aligned} \tag{17}$$

Again, by setting $S = 1$, we get

$$\frac{\partial}{\partial t}\left(\frac{\partial^2 G}{\partial S^2}\right)\Bigg|_{S=1} = -4k_1 \left.\frac{\partial^3 G}{\partial S^3}\right|_{S=1} - k_1 \left.\frac{\partial^4 G}{\partial S^4}\right|_{S=1} - k_1 \left.\frac{\partial^4 G}{\partial S^4}\right|_{S=1}$$
$$+4k_2 \left.\frac{\partial^2 G}{\partial S^2}\right|_{S=1} + k_2 \left.\frac{\partial^3 G}{\partial S^3}\right|_{S=1} + k_2 \left.\frac{\partial^3 G}{\partial S^3}\right|_{S=1} - k_3 \left.\frac{\partial^2 G}{\partial S^2}\right|_{S=1}$$
$$-k_3 \left.\frac{\partial^2 G}{\partial S^2}\right|_{S=1} + k_4 \left.\frac{\partial G}{\partial S}\right|_{S=1} + k_4 \left.\frac{\partial G}{\partial S}\right|_{S=1}. \tag{18}$$

Or from Eq. (11)

$$\frac{\partial\{X^2\}}{\partial t} = -2k_1\{X^4\} + (-4k_1 + 2k_2)\{X^3\} + (4k_2 - 2k_3)\{X^2\} + 2k_4\{X\}. \tag{19}$$

This is the second-order factorial moment equation. All higher order moments can be constructed with this recursive algorithm.

For nonlinear reactions this set of equations must, in principle, be extended to infinite order. The reason becomes evident by looking at Eq. (16), where the first moment depends on the second and the third. In Eq. (19) it is evident that $\{X^2\}$ depends on $\{X^3\}$ and $\{X^4\}$. Similarly, in the equation for $\{X^M\}$, where M is an arbitrarily large, yet finite, moment, we would find terms like $\{X^{(M+1)}\}$, $\{X^{(M+2)}\}$, etc.

We can concisely write the moment equations as follows:

$$\frac{\partial \boldsymbol{\mu}}{\partial t} = A\boldsymbol{\mu} + A'\boldsymbol{\mu'}, \tag{20}$$

where $\boldsymbol{\mu}$ is the vector of lower-order moments

$$\boldsymbol{\mu} = \left[1\ \{X\}\ \{X^2\}\ \{X^3\}\ \{X^4\}\ \{X^5\} \ldots \{X^M\}\right]^T \tag{21}$$

and $\boldsymbol{\mu'}$ is the vector of higher-order moments

$$\boldsymbol{\mu'} = \left[\{X^{(M+1)}\}\ \{X^{(M+2)}\} \ldots\right]^T. \tag{22}$$

A and A' are matrices constructed with coefficients from moment equations.

The major challenge with solving the moment Eq. (20) is the result of non-zero elements present in matrix A'. The set never closes for nonlinear reactions, requiring an infinite number of moment equations. Elsewhere, we explored numerous, previously proposed closure schemes [12].

One insight to move past this impasse is that for reaction networks with finite state space we may anticipate a probability distribution that can be adequately

described by a finite number of moments. Note this argument is void of mathematical rigor. The insight is purely based on physical intuition. Let us accept it for now with the understanding that it may or may not serve us well, depending on the reaction network.

Empirically, we find that the probability distribution of the Schlögl model can indeed be described within reasonable numerical accuracy with up to 12 lower-order moments. For the sake of brevity, we will present the algorithm assuming that six moments suffice. We also focus only on steady state probability distributions. We have discussed non-steady state solutions of moment equations in [1].

At steady state, moment equations become

$$\mathbf{0} = A\boldsymbol{\mu} + A'\boldsymbol{\mu'}, \tag{23}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\mu'}$ include the stationary probability distribution moments.

For the Schlögl model with up to six lower order moments (excluding the zero order moment) we have

$$\begin{aligned} \boldsymbol{\mu} &= \left[1\ \{X\}\ \{X^2\}\ \{X^3\}\ \{X^4\}\ \{X^5\}\ \{X^6\}\right]^T \\ \boldsymbol{\mu'} &= \left[\{X^7\}\ \{X^8\}\right]^T . \end{aligned} \tag{24}$$

The zero order moment is always 1, since $\{X^0\} = \sum_{x=0}^{\infty} \frac{x!}{(x-0)!} P(x,t) = \sum_{x=0}^{\infty} P(x,t) = 1$.

Matrices A and A' are presented in Eqs. (25) and (26), respectively:

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ k_4 & -k_3 & k_2 & -k_1 & 0 & 0 & 0 \\ 0 & 2k_4 & 4k_2 - 2k_3 & 2k_2 - 4k_1 & -2k_1 & 0 & 0 \\ 0 & 0 & 6k_2 + 3k_4 & 12k_2 - 6k_1 - 3k_3 & 3k_2 - 12k_1 & -3k_1 & 0 \\ 0 & 0 & 0 & 24k_2 + 4k_4 & 24k_2 - 24k_1 - 4k_3 & 4k_2 - 24k_1 & -4k_1 \\ 0 & 0 & 0 & 0 & 60k_2 + 5k_4 & 40k_2 - 60k_1 - 5k_3 & 5k_2 - 40k_1 \\ 0 & 0 & 0 & 0 & 0 & 120k_2 + 6k_4 & 60k_2 - 120k_1 - 6k_3 \end{bmatrix} \tag{25}$$

$$A' = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ -5k_1 & 0 \\ 6k_2 - 60k_1 & -6k_1 \end{bmatrix} . \tag{26}$$

In principle, any CME can be transformed into moment equations, through Z-transform. At this point instead of the CME we have a finite set of ordinary differential equations, albeit one that is not solvable because of the non-zero elements in A'.

3 ZI-Closure Scheme

3.1 *Solving Moment Equations by Maximizing the Information Entropy*

In order to resolve the numerical closure issues of solving moment equations, we proposed the zero-information closure scheme [1]. We assume that all information necessary to build the probability distribution described by the CME is contained within a finite number of lower-order moments. In that case, higher-order moments add no information to the reconstruction of the probability distribution and may be obtained from the maximization of information entropy.

Using Shannon's definition [21], information entropy is given by

$$H = -\sum_{x=0}^{\infty} P(x) \ln P(x). \tag{27}$$

Again, x can take all possible values of the state variable X.

Assuming that the first M lower-order moments (not including the zero-order one) are known, the entropy defined in Eq. (27) must be maximized with respect to these constraints. This can be accomplished with the method of Lagrange multipliers:

$$\Lambda = H - \sum_{i=0}^{M} \lambda_i \left[\sum_{x=0}^{\infty} \frac{x!}{(x-i)!} P(x) - \{X^i\} \right], \tag{28}$$

where Λ is the Lagrangian and λ_i is the Lagrange multiplier associated with the lower-order moment $\{X^i\}$.

Since the entropy is maximum, for every value $x = y$

$$\frac{\partial \Lambda}{\partial P(y)} = 0 \tag{29}$$

or from Eq. (28)

$$\frac{\partial H}{\partial P(y)} - \frac{\partial}{\partial P(y)} \left\{ \sum_{i=0}^{M} \lambda_i \left[\sum_{x=0}^{\infty} \frac{x!}{(x-i)!} P(x) - \{X^i\} \right] \right\} = 0. \tag{30}$$

From Eq. (27), we obtain

$$\frac{\partial H}{\partial P(y)} = \frac{\partial}{\partial P(y)}\left[-\sum_{x=0}^{\infty} P(x)\ln P(x)\right] = -\frac{\partial}{\partial P(y)}\left[P(y)\ln P(y)\right] = -\ln P(y) - 1. \tag{31}$$

Additionally,

$$\begin{aligned}
&\frac{\partial}{\partial P(y)}\left\{\sum_{i=0}^{M}\lambda_i\left[\sum_{x=0}^{\infty}\frac{x!}{(x-i)!}P(x) - \{X^i\}\right]\right\} \\
&\quad = \sum_{i=0}^{M}\lambda_i\left\{\frac{\partial}{\partial P(y)}\left[\sum_{x=0}^{\infty}\frac{x!}{(x-i)!}P(x)\right] - \frac{\partial\{X^i\}}{\partial P(y)}\right\} \\
&\quad = \sum_{i=0}^{M}\lambda_i\frac{y!}{(y-i)!}\frac{\partial P(y)}{\partial P(y)} = \sum_{i=0}^{M}\lambda_i\frac{y!}{(y-i)!}.
\end{aligned} \tag{32}$$

Combining Eqs. (30)–(32) we conclude

$$P_H(y) = \exp\left[-1 - \sum_{i=0}^{M}\lambda_i\frac{y!}{(y-i)!}\right]. \tag{33}$$

Equation (33) gives the stationary probability distribution for the maximum entropy solely in terms of the Lagrange multipliers. The subscript H emphasizes that the distribution is given for the maximum information entropy distribution.

Through the expression of the probability distribution for maximum entropy, both the lower-order and higher-order moments can now be related to a set of $M+1$ Lagrange multipliers. The following holds for every moment of the stationary probability distribution

$$\{X^m\}_H = \sum_{x=0}^{\infty}\frac{x!}{(x-m)!}P_H(x) = \sum_{x=0}^{\infty}\frac{x!}{(x-m)!}\exp\left[-1 - \sum_{i=0}^{M}\lambda_i\frac{x!}{(x-i)!}\right]. \tag{34}$$

For example, the third moment of the Schlögl model example for up to $M = 6$ lower-order moments is given by

$$\begin{aligned}
\{X^3\}_H = \sum_{x=0}^{\infty}\frac{x!}{(x-3)!}\exp\left[-1 - \sum_{i=0}^{6}\lambda_i\frac{x!}{(x-i)!}\right] &= \sum_{x=0}^{\infty} x\,(x-1)\,(x-2) \\
&\times \exp\left[-1 - \lambda_0 - \lambda_1 x - \lambda_2 x\,(x-1) - \lambda_3 x\,(x-1)\,(x-2)\right. \\
&-\lambda_4 x\,(x-1)\,(x-2)\,(x-3) - \lambda_5 x\,(x-1)\,(x-2)\,(x-3)\,(x-4) \\
&\left.-\lambda_6 x\,(x-1)\,(x-2)\,(x-3)\,(x-4)\,(x-5)\right].
\end{aligned} \tag{35}$$

Based on Eq. (34), vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu'}$ of the Schlögl model for up to six lower-order moments can now be written only in terms of the seven Lagrange multipliers:

$$\boldsymbol{\mu}_H = \begin{bmatrix} 1 \\ \sum_{x=0}^{\infty} \frac{x!}{(x-1)!} \exp\left[-1 - \sum_{i=0}^{6} \lambda_i \frac{x!}{(x-i)!}\right] \\ \sum_{x=0}^{\infty} \frac{x!}{(x-2)!} \exp\left[-1 - \sum_{i=0}^{6} \lambda_i \frac{x!}{(x-i)!}\right] \\ \sum_{x=0}^{\infty} \frac{x!}{(x-3)!} \exp\left[-1 - \sum_{i=0}^{6} \lambda_i \frac{x!}{(x-i)!}\right] \\ \sum_{x=0}^{\infty} \frac{x!}{(x-4)!} \exp\left[-1 - \sum_{i=0}^{6} \lambda_i \frac{x!}{(x-i)!}\right] \\ \sum_{x=0}^{\infty} \frac{x!}{(x-5)!} \exp\left[-1 - \sum_{i=0}^{6} \lambda_i \frac{x!}{(x-i)!}\right] \\ \sum_{x=0}^{\infty} \frac{x!}{(x-6)!} \exp\left[-1 - \sum_{i=0}^{6} \lambda_i \frac{x!}{(x-i)!}\right] \end{bmatrix} \tag{36}$$

$$\boldsymbol{\mu'}_H = \begin{bmatrix} \sum_{x=0}^{\infty} \frac{x!}{(x-7)!} \exp\left[-1 - \sum_{i=0}^{6} \lambda_i \frac{x!}{(x-i)!}\right] \\ \sum_{x=0}^{\infty} \frac{x!}{(x-8)!} \exp\left[-1 - \sum_{i=0}^{6} \lambda_i \frac{x!}{(x-i)!}\right] \end{bmatrix}. \tag{37}$$

Matrices A and A' still have the same form [Eqs. (25) and (26)]. The difference is that now Eq. (23) depends only on a set of seven Lagrange multipliers. Given that there is a set of six coupled equation as well as one normalization constraint with seven unknown parameters, then the Lagrange multipliers, and hence the probability distribution, can be computed using a root-finding numerical method such as Newton–Raphson.

3.2 *Newton–Raphson Algorithm for Finding the Lagrange Multipliers*

The steady state moment equation [Eq. (23)] depends only on Lagrange multipliers [Eqs. (36) and (37)] and the kinetic constants of the system [Eqs. (25) and (26)]. Since, the system is closed, i.e. it has the same number of unknowns and equations, a root-finding method can be equipped in order to find the steady state Lagrange multipliers.

A proposed algorithm based on the Newton–Raphson method can be

1. Set the necessary number of lower-order moments (M). The value of M can affect the accuracy of the solution since by setting a specific number of lower-order moments, we assume that the rest of them have little to offer in the solution. For the Schlögl model example $M = 12$ is sufficient [1].

2. Calculate matrices A and A' through Eqs. (25) and (26). These matrices depend only on the kinetic constants of the system.
3. Introduce an initial guess for the Lagrange multipliers of the system $\boldsymbol{\lambda} = [\lambda_0, \lambda_1, \ldots, \lambda_M]$. Usually this guess is either a vector of zeros (which means that a uniform distribution is the initial guess) or the Lagrange multipliers generated from a system close to the one we are trying to solve.
4. Calculate the moment vectors $\boldsymbol{\mu}_{\boldsymbol{\lambda}}$ and $\boldsymbol{\mu'}_{\boldsymbol{\lambda}}$ based on Eqs. (36), (37) and the current value of the Lagrange multipliers $\boldsymbol{\lambda}$. The subscript $\boldsymbol{\lambda}$ is employed to emphasize that the vectors have values based on the current guess of Lagrange multipliers.
5. Calculate the value of the moment equations for the current guess of moments

$$\Delta \boldsymbol{\mu}_{\boldsymbol{\lambda}} = A\boldsymbol{\mu}_{\boldsymbol{\lambda}} + A'\boldsymbol{\mu'}_{\boldsymbol{\lambda}}. \tag{38}$$

6. Calculate the Euclidean norm of vector $\Delta \boldsymbol{\mu}_{\boldsymbol{\lambda}}$

$$\epsilon = \sqrt{\Delta \boldsymbol{\mu}_{\boldsymbol{\lambda}}^T \Delta \boldsymbol{\mu}_{\boldsymbol{\lambda}}}. \tag{39}$$

7. If ϵ is less than the wanted accuracy threshold proceed to (8). Otherwise:
   - Calculate the Jacobian of the system:

$$J = \frac{\partial \Delta \boldsymbol{\mu}_{\boldsymbol{\lambda}}}{\partial \boldsymbol{\lambda}} = A\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\lambda}}}{\partial \boldsymbol{\lambda}} + A'\frac{\partial \boldsymbol{\mu'}_{\boldsymbol{\lambda}}}{\partial \boldsymbol{\lambda}}. \tag{40}$$

   - Based on Newton–Raphson method, calculate a first order Taylor expansion:

$$\Delta \boldsymbol{\mu}_{\boldsymbol{\lambda}} = J\Delta \boldsymbol{\lambda}. \tag{41}$$

   - Create an approximate step $\Delta \boldsymbol{\lambda}$ for the Lagrange multipliers vector $\boldsymbol{\lambda}$:

$$\Delta \boldsymbol{\lambda} = J^{-1}\Delta \boldsymbol{\mu}_{\boldsymbol{\lambda}}. \tag{42}$$

     where J^{-1} is the inverse of the Jacobian matrix J.
   - Calculate the new set of Lagrange multipliers: $\boldsymbol{\lambda} = \boldsymbol{\lambda} + \Delta \boldsymbol{\lambda}$.
   - Return to (4).
8. The final solution is $\boldsymbol{\lambda}$ and the probability distribution is calculated based on $\boldsymbol{\lambda}$ and Eq. (33).

3.3 Some Results for the Schlögl Model

Using this algorithm we may solve stochastic reaction networks, including the Schlögl model as previously reported [1]. For a specific set of kinetic constants, the stationary probability distribution calculated with ZI -closure scheme, for a 12th order closure ($M = 12$), is presented in Fig. 1. As the figure shows, ZI-closure accurately computes the steady state distribution compared to results from Gillespie's stochastic simulation algorithm (SSA) [15] computed using our SSA solver [22].

Table 1 also presents the first eight moments calculated with ZI-closure scheme and SSA, as well as the corresponding Lagrange multipliers calculated for ZI-closure. The moment values calculated of the ZI-closure scheme are close to the ones of SSA (forth column of Table 1).

Finally, Fig. 2 shows how the solution of ZI-closure scheme is affected by the chosen order of the lower-order moments M. The figure demonstrated that at least ten moments are needed to accurately reconstruct the probability distribution. After a specific value of lower-order moments ($M = 10$), the method converges to the same probability distribution.

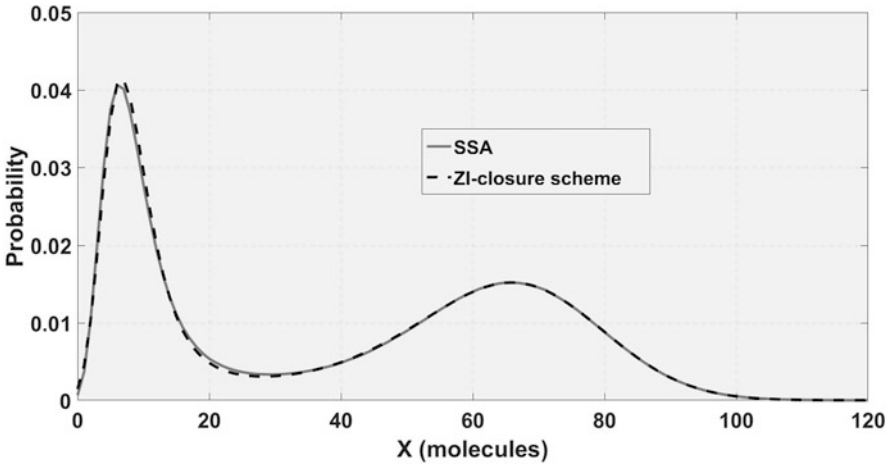


Fig. 1 The figure compares the solution of ZI-closure scheme for 12th order closure and the solution of SSA for at least 500,000 trajectories. The solution of ZI-closure schemes matches the one of SSA. For this figure, Table 1 and Fig. 2 the kinetic constants are $(k_1, k_2, k_3, k_4) = (0.0015, 0.15, 3.5, 20)$

Table 1 The second and third columns of the table present the values of the moments of the stationary probability distribution of SSA and ZI-closure scheme, respectively

	SSA	ZI-closure scheme	Difference (%)	Lagrange multiplier
First moment	40	40	6.0×10^{-6}	-1.1
Second moment	2.4×10^{3}	2.4×10^{3}	4.5×10^{-6}	1.5×10^{-1}
Third moment	1.6×10^{5}	1.6×10^{5}	1.0×10^{-5}	-8.5×10^{-3}
Fourth moment	1.1×10^{7}	1.1×10^{7}	1.2×10^{-5}	2.9×10^{-4}
Fifth moment	7.9×10^{8}	7.9×10^{8}	1.3×10^{-5}	-6.2×10^{-6}
Sixth moment	5.7×10^{10}	5.7×10^{10}	1.3×10^{-5}	9.1×10^{-8}
Seventh moment	4.2×10^{12}	4.2×10^{12}	1.3×10^{-5}	-8.9×10^{-10}
Eighth moment	3.1×10^{14}	3.1×10^{14}	1.2×10^{-5}	5.5×10^{-12}

The rows of the table indicate the order of the moment. The percentage difference between the moments calculated with the two methods relative to SSA results is located at the fourth column. The fifth column presents the corresponding Lagrange multipliers calculated with ZI-closure

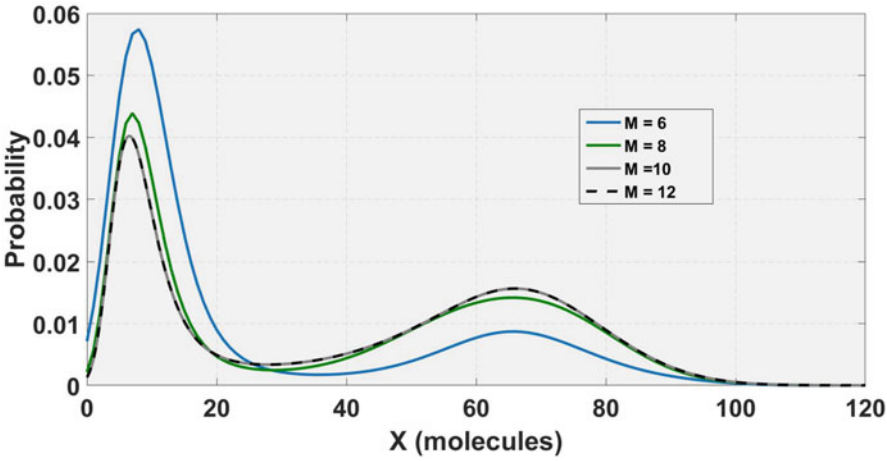


Fig. 2 The stationary probability distribution of ZI-closure scheme for different closure orders, M

4 Concluding Remarks

In this chapter we explored in detail the use of factorial moments and the zero-information closure scheme for solving CMEs. Numerous chemical reaction networks with diverse dynamics have been numerically explored (e.g., Michaelis–Menten system, closed dimerization reaction, cyclic chains, etc.) and can be found in [1, 12]. Here we have presented the Schlögl model, a simple, single-component system.

There are numerous challenges we face when using ZI-closure. An important one is that ZI-closure is currently limited to one- or two-dimensional systems; that is, networks with only up to two chemical degrees of freedom. This is a practical,

numerical implementation challenge, a result of the exploding phase space size for multidimensional problems.

Furthermore, for more than two reacting species, the number of moment equations quickly increases to impractical sizes. This is because even for low moment order closure scheme, the moments include cross variable terms, e.g., X^2, Y^2 and also XY, or X^3, Y^3, X^2Y and XY^2.

Another important limitation is that there is currently no way to know a priori how many moments suffice for a numerically accurate reconstruction of the probability distribution. Furthermore, the reconstruction itself becomes a progressively more challenging task as the number of moments increases.

ZI-closure is certainly not an efficient method to compute the dynamics of probability distribution in non-steady state systems. This is because, at each time step an entropy maximization step must be implemented. Although Lagrange multiplier values from the previous time step can be used as an initial condition, the numerical implementation is slow.

Nevertheless, with ZI-closure there is now a tool to compute steady state probability distributions, without having to resort to stochastic simulations in time. This is an important advantage of the method compared to other stochastic simulation methods.

Furthermore, with ZI-closure the tools become available to explore the stability of reaction network steady states and the sensitivity of the behavior to changes in system parameters [12, 23], paving another avenue for investigating stochastic reaction networks.

The entropy of a chemical reaction network at non-equilibrium steady state (NESS) is at the heart of ZI-closure. We are wondering whether NESS are established when the entropy is maximum. We are currently pursuing a research program to clarify the physico-chemical implications of the postulate of maximum entropy imposed by ZI-closure.

There is much work left on developing ZI-closure. We hope to engage the community in a dialogue on what is important in stochastic dynamics simulations and on how to improve CME solvers.

Acknowledgements This work was supported by a grant from the National Institutes of Health (GM111358) and a grant from the National Science Foundation (CBET-1412283). This work utilized the high-performance computational resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575. Support from the University of Minnesota Digital Technology Center, from the Minnesota Supercomputing Institute (MSI), and from CAPES—Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil is gratefully acknowledged. This work was partially completed Spring 2016, when YNK was Visiting Scholar at the Isaac Newton Institute of Mathematical Sciences at the University of Cambridge, attending the programme Stochastic Dynamical Systems in Biology.

References

1. P. Smadbeck, Y.N. Kaznessis, A closure scheme for chemical master equations. Proc. Natl. Acad. Sci. U. S. A. **110**, 14261–14265 (2013)
2. K.R. Popper, *The Open Universe: An Argument for Indeterminism* (Cambridge, Routledge, 1982), p. xix
3. W. James, *The Dilemma of Determinism*. The Will to Believe (New York, Dover, 1956)
4. I. Prigogine, *The End of Certainty: Time, Chaos, and the New Laws of Nature* (Free Press, New York, 1997)
5. D.A. McQuarrie, Stochastic approach to chemical kinetics. J. Appl. Probab. **4**, 413–478 (1967)
6. I. Oppenheim, K.E. Shuler, Master equations and Markov processes. Phys. Rev. **138**, B1007–B1011 (1965)
7. N.G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, Revised and enlarged edition (Elsevier, Amsterdam, 2004)
8. D.T. Gillespie, A rigorous derivation of the chemical master equation. Physica A **188**, 404–425 (1992)
9. D.T. Gillespie, Stochastic simulation of chemical kinetics. Ann. Rev. Phys. Chem. **58**, 35–55 (2007)
10. D.T Gillespie, A general method for numerically simulating the stochastic time evolution of coupled reactions. J. Comput. Phys. **22**, 403–434 (1976)
11. Y. Kaznessis, Multi-scale models for gene network engineering. Chem. Eng. Sci. **61**, 940–953 (2006)
12. P.H. Constantino, M. Vlysidis, P. Smadbeck, Y.N. Kaznessis, Modeling stochasticity in biochemical reaction networks. J. Phys. D Appl. Phys. **49**, 093001 (2016)
13. V. Sotiropoulos, Y.N. Kaznessis, Analytical derivation of moment equations in stochastic chemical kinetics. Chem. Eng. Sci. **66**, 268–277 (2010)
14. P. Smadbeck, Y.N. Kaznessis, Efficient moment matrix generation for arbitrary chemical networks. Chem. Eng. Sci. **84**, 612–618 (2012)
15. C.S. Gillespie, Moment-closure approximations for mass-action models. IET Syst. Biol. **3**, 52–58 (2009)
16. E.T. Jaynes, Information theory and statistical mechanics. Phys. Rev. **106**, 620–630 (1957)
17. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–659 (1948)
18. F. Schlögl, On thermodynamics near a steady state. Z. Phys. **248**, 446–458 (1971)
19. D.T. Gillespie, *Markov Processes, An Introduction for Physical Scientists* (Academic, Cambridge, 1992)
20. M.H. DeGroot, M.H. Schervish, *Probability and Statistics*, 4th edn. (Pearson, Cambridge, 2012)
21. J.N. Kapur, *Maximum-Entropy Models in Science and Engineering*, 1st edn. (Wiley, New York, 1989)
22. A.D. Hill, J.R. Tomshine, E.M. Weeding, V. Sotiropoulos, Y.N. Kaznessis, SynBioSS: the synthetic biology modeling suite. Bioinformatics **24**, 2551–2553 (2008)
23. P. Smadbeck, Y.N. Kaznessis, On a theory of stability for nonlinear stochastic chemical reaction networks. J. Chem. Phys. **142**, 184101 (2015)

Deterministic and Stochastic Becker–Döring Equations: Past and Recent Mathematical Developments

E. Hingant and R. Yvinec

1 Introduction

The Becker–Döring (BD) equations go back to the seminal work "Kinetic treatment of nucleation in supersaturated vapors" by Becker and Döring (1935), which gave rise to the name of the model. Later on, Burton (1977) popularized the use of such equations to study condensations phenomena at different pressures. Since then, applications of this model range from physics, chemistry to biology. Recently, the book edited by Schmelzer (2005) makes an inventory of several applications of nucleation and phase transition theory. Let us also point out recent applications of Becker–Döring or related coagulation-fragmentation models in biology, specifically to protein aggregation in neurodegenerative diseases, e.g. the works by Linse and Linse (2011), Prigent et al. (2012), Alvarez-Martinez et al. (2011), Budrikis et al. (2014), Eden et al. (2015), Davis and Sindi (2016), Eugene et al. (2016) and Doumic et al. (2016). Also Hu and Othmer (2011) worked on polymerization of actin filaments, Hoze and Holcman (2014; 2015) on assembly of virus capsids, Bressloff (2016) on vesicular transport, and Hoze and Holcman (2012) for telomere clustering.

In its survey, Slemrod (2000) said the BD equations "provide perhaps the simplest kinetic model to describe a number of issues in the dynamics of phase transitions." This is maybe one of the reasons these equations received lots of

E. Hingant
Departamento de Matemática, Universidad del Bío-Bío, Concepción, Chile
e-mail: ehingant@ubiobio.cl

R. Yvinec (✉)
CR2 INRA, UMR85 Physiologie de la Reproduction et des Comportements, F-37380 Nouzilly, France
e-mail: romain.yvinec@tours.inra.fr

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_9

attention from many mathematicians. But despite their simplicity, these equations are rich and difficult. Our intention here is: on one hand, to complete the review by Slemrod with new results; and on the other hand, to give a parallel with the stochastic version of these equations, which reveals a lot of new interesting problems. We also mention the review by Wattis (2006) which contains many qualitative and exact properties of the solutions in the deterministic context, and the pedagogical notes by Penrose (1995). But, few stochastic reviews of the BD model are available, we can only mention the seminal work by Aldous (1999) which treats the so-called Smoluchowsky coagulation equations.

The model consists in describing the repartition of clusters by their size $i \geq 1$, i.e. the number of particles that composed them. Clusters belong to a "solvent" in much smaller proportion and are assumed to be spatially homogeneously distributed. Along their motion, clusters give rise to two types of reactions, namely the *Becker–Döring rules*:

1. A cluster of size 1, commonly called *monomer* or *elementary particle*, may encounter a cluster of size $i \geq 1$ to coalesce and give rise to a cluster of size $i + 1$.
2. A cluster of size $i \geq 2$ may release spontaneously a monomer resulting in a cluster of size $i - 1$ and a cluster of size 1.

These can be summarized by the set of kinetic reactions, for each $i \geq 1$,

$$C_1 + C_i \underset{b_{i+1}}{\overset{a_i}{\rightleftharpoons}} C_{i+1}\,, \tag{1}$$

where C_i denotes clusters consisting of i particles. Coefficients a_i and b_{i+1} stand, respectively, for the rate of aggregation and fragmentation. These may depend on the size of clusters involved in the reactions and typical coefficients are derived by Penrose (1997) and Niethammer (2003):

$$a_i = i^\alpha\,, \quad b_{i+1} = a_{i+1}\left(z_s + \frac{q}{(i+1)^\gamma}\right)\,, \qquad i \geq 1\,. \tag{2}$$

for $0 \leq \alpha < 1$, $z_s > 0$, $q > 0$ and $0 < \gamma < 1$. This choice is in agreement with original derivation where $a_i \approx i^{2/3}$, $b_i \approx a_i \exp(Gi^{-1/3})$. In particular, the diffusion-limited case of monomers clustering into sphere corresponds to $\alpha = 1/3$, $\gamma = 1/3$ in 3D and to $\alpha = 0$, $\gamma = 1/2$ in 2D, while the interface-reaction-limited case corresponds to $\alpha = 2/3$, $\gamma = 1/3$ in 3D and $\alpha = 1/2$, $\gamma = 1/2$ in 2D. We refer also to Penrose and Buhagiar (1983) for a method on deriving coefficients. Note that all along the survey we assume the natural hypothesis a_i and b_{i+1} are non-negative for all $i \geq 1$, without referring to this again.

In its mean-field version, or deterministic, the BD model is an infinite set of ordinary differential equations for the time evolution of each concentrations (numbers per unit of volume) of clusters made of i particles. In its stochastic version, the BD model is a continuous time Markov chain, on a finite state space. We divide the remainder of this survey into two parts for the respective versions.

2 Deterministic Mean-Field Theory

The general formulation of the deterministic Becker–Döring equations, as studied today, seems to go back to Burton (1977) and was popularized among mathematicians by Penrose and Lebowitz (1979) (indeed, the equations studied in the original work by Becker and Döring (1935) slightly differ, see comment later on). It assumes the system behaves homogeneously in space with a high number of clusters, and considers concentrations $c_i(t)$ (unit per volume) of clusters with size $i \geq 1$ at time $t \geq 0$. It deals with classical law of chemistry (Law of Mass Action), the coagulation is considered as a second order reaction while the fragmentation is a first-order (linear) reaction. The flux associated with the kinetic scheme (1) is thus given, for each $i \geq 1$, by

$$J_i = a_i c_1 c_i - b_{i+1} c_{i+1} \,. \tag{3}$$

Considering all the fluxes involved in the variations of the concentration of each c_i entails the infinite system of differential equations, namely the Becker–Döring equations:

$$\frac{d}{dt} c_1 = -J_1 - \sum_{i \geq 1} J_i \,, \tag{4}$$

$$\frac{d}{dt} c_i = J_{i-1} - J_i \,, \tag{5}$$

for every $i \geq 2$. The system considered here has no source nor sink. Consequently, for the total amount of monomers, we should have, for all $t \geq 0$,

$$\sum_{i \geq 1} i c_i(t) = \rho \,, \tag{6}$$

where ρ is a constant, called through the survey: *mass of the system*. Formal computations on the solution of the system (4)–(5), interverting infinite sum, lead to this statement. Remark, the constant ρ is entirely determined by the initial condition given at time $t = 0$. In this section we try to expound the main theory around these equations. In particular, we exclude many variants such as the original constant monomer formulation, which is then an infinite linear system (e.g., Penrose 1989, Kreer 1993, or King and Wattis 2002), the finite-dimensional truncated system (e.g., Duncan and Soheili 2001 or Duncan and Dunwell 2002), generalization such as micelles formation (e.g., Coveney and Wattis 1996) or including space with cluster diffusion (e.g., Laurençot and Wrzosek 1998) or lattice models (e.g., Penrose and Buhagiar 1983).

We separated this section between well-posedness, long-time behavior, scaling limit, and some time-dependent properties.

2.1 Well-Posedness

The first *general* result on existence and uniqueness on Becker–Döring equations is due to Ball et al. (1986) which really starts the mathematical analysis of BD equations. The authors state many of the fundamental properties of the solutions belonging to the Banach space

$$X^+ := \left\{ x \subset \mathbb{R}_+^{\mathbb{N}} \, : \, \sum_{i\geq 1} i x_i < +\infty \right\} ,$$

which arises naturally in view of the balance of mass Eq. (6). We recall first the notion of solutions to BD equations.

Definition 1 Let $T \in (0, +\infty]$ and $c^{\text{in}} \in X^+$. A solution to the Becker–Döring equations (4)–(5) on $[0, T)$ with initial data c^{in} is a function $c : [0, T) \to X^+$ which writes $c := (c_i)_{i\geq 1}$ and such that: $\sup_{t\in[0,T)} \|c(t)\|_X < +\infty$; for all $t \in [0, T)$, we have $\sum_{i\geq 1} a_i c_i \in L^1(0, t)$ and $\sum_{i\geq 2} b_i c_i \in L^1(0, t)$; Eqs. (4)–(5) hold almost every $t \in [0, T)$ and $c(0) = c^{\text{in}}$.

One of the fundamental facts, proved by Ball et al. (1986), is that any solution to the BD equations satisfies the balance of mass Eq. (6) at all finite time (Corollary 2.6). In particular, any solution to the BD equations avoids the so-called gelation phenomenon (in finite time) which can occur in general coagulation-fragmentation equations (e.g., Escobedo et al. 2003). Ball et al. (1986) also proved propagation of moments (Theorem 2.2) and regularity properties of the solutions (Theorem 3.2). Finally, they state a general existence result for sublinear coagulation rate and uniqueness with an extra-moment on the initial condition (see below Theorem 1). In short, the work by Ball et al. (1986) covered the essential properties of BD equations, build the foundations for the analysis of BD equations, and should be a companion for whom want to work with.

We go back to well-posedness, for which Laurençot and Mischler (2002) complement the result in Ball et al. (1986), proving the uniqueness without extra condition on the initial data but assuming a growth condition on the fragmentation rate, viz. there exists a constant $K > 0$ such that, for every $i \geq 2$,

$$a_i - a_{i-1} \leq K , \quad b_i - b_{i+1} \leq K . \tag{7}$$

We summarize these results in the following theorem.

Theorem 1 (Well-Posedness, Ball et al. 1986, Laurençot and Mischler 2002) *Let $c^{\text{in}} \in X^+$. Assume alternatively either (a) $a_i = O(i)$ and $\sum_{i\geq 1} i^2 c_i^{\text{in}} < +\infty$, or (b) the growth condition (7). The Becker–Döring equations (4)–(5) have a unique solution c on $[0, +\infty)$ associated with the initial data c^{in}. Moreover, for all $t \geq 0$,*

$$\sum_{i\geq 1} i c_i(t) = \sum_{i\geq 1} i c_i^{\text{in}} .$$

In fact the uniqueness in Ball et al. (1986) is slightly more subtle, see their Theorem 3.6. Also, they proved that $a_i = O(i)$ is almost optimal. Indeed, their Theorem 2.7 states: if $\lim_{i\to\infty} a_i/i = +\infty$ and $\lim_{i\to\infty} b_{i+1}/a_i < +\infty$, then for some initial condition (with relatively *fat* tail) still belonging to X^+ the BD system has no-solution. This suggests that, for super-linear coagulation rate, we cannot hope existence for a large class of initial data without a sufficient control on the fragmentation rate. Since mass is preserved, fragmentation should balance the formation of "big" clusters. It seems very few results exist for such class of coefficients, except Wattis et al. (2004) who considered exponential coefficients.

Finally, we mention that a proof of existence to the BD equations is self-contained in the nice proof by Laurençot (2002) for a more general model (discrete coagulation with multiple fragmentation). It relies, as for the proof given by Ball et al. (1986), on a truncated system up to a size N and compactness arguments to obtain the limit $N \to +\infty$. But here Laurençot (2002) took advantage of the propagation of super-linear moments and a De La Vallé Poussin lemma to prove compactness.

2.2 *Long-Time Behavior*

The long-time behavior of the BD system brings some of its most interesting properties, and we will see this is still under active research. Through this section we will always assume that both a_i and b_{i+1} are positive for each $i \geq 1$. This avoids many *pathological* cases, in some sense, if one of them cancels it "breaks the communication" between clusters in one side or another. Nonetheless, we mention the interesting cases (not detailed here) where either $a_i = 0$ or $b_{i+1} = 0$, for every $i \geq 1$, which have been treated again by Ball et al. (1986)! We start with a subsection which deals with convergence to equilibrium. Then, we will see the most recent results on the exponential stability of the equilibrium.

2.2.1 Convergence to Equilibrium

The equilibrium candidates, at plural, of the BD equations are obtained by canceling the fluxes J_i, for each $i \geq 1$, as defined in Eq. (3). After straightforward manipulation of the fluxes, the candidates form a one-parameter family, indexed by a variable $z \geq 0$, and are given by the expressions

$$\bar{c}_i(z) = Q_i z^i\,, \quad \text{where} \quad Q_i = \frac{a_1 a_2 \cdots a_{i-1}}{b_2 b_3 \cdots b_i}\,, \tag{8}$$

for every $i \geq 1$, with the convention $Q_1 = 1$. For example, the case related to Eq. (2) gives e.g., Niethammer (2003), for large i,

$$Q_i \approx \frac{C}{i^{\alpha} z_s^{(i-1)}} \exp\left(-\frac{q}{(1-\gamma)z_s} i^{1-\gamma}(1 + O(i^{-\gamma}))\right) .$$

To find the right equilibrium, which reduces to find the value of $\bar{c}_1 = z$, one should use the balance of mass Eq. (6) which we know to be satisfied at any finite time. Hence, this leads us to consider the power series, given by the mass of the equilibrium candidates,

$$\sum_{i \geq 1} i Q_i z^i ,$$

which radius of convergence is denoted by z_s. Such radius is obtained from the rates functions since the Cauchy–Hadamard theorem says $1/z_s := \limsup_{i \to \infty} Q_i^{1/i}$ [note that this definition is consistent with the z_s that was used in the example given by Eq. (2)]. This becomes the heart of the existence of a critical mass in BD equations since the values taken by the series may not define a bijection from $[0, z_s)$ into $[0, +\infty)$. Indeed, set ρ_s be the upper value taken by the series on $\{z < z_s\}$. It can occur that ρ_s is finite, in which case we already know that there is no equilibrium candidate with mass $\rho > \rho_s$. Hence, this leads to a dichotomy in the long-time behavior of the BD equations whether or not the mass of the solution considered is less than ρ_s, named the *critical mass*. We may refer to sub-critical solution when the mass $\rho < \rho_s$, critical solution when $\rho = \rho_s$ and super-critical solution when $\rho > \rho_s$.

The Becker–Döring equations are part of the kinetic equations. The latter have a long story, led by the celebrated Boltzmann equations, which are of course completely out of the scope of this paper, maybe the reader could refer to Cercignani (1990). The key concept in these equations is the *entropy* (sometimes called energy) which, in mathematical words, is a Lyapunov functional and governs the trend to equilibrium. Namely, the entropy arising in BD equations is given by the expression

$$H(c) = \sum_{i \geq 1} c_i \left(\ln\left(\frac{c_i}{Q_i}\right) - 1\right) .$$

This is because, formally, the H decreases along the solutions c (and is bounded from below), as

$$\frac{d}{dt} H(c(t)) = -D(c(t)) , \tag{9}$$

where the *dissipation* is

$$D(c) := \sum_{i\geq 1} (a_i c_1 c_i - b_{i+1} c_{i+1}) \left(\ln(a_i c_1 c_i) - \ln(b_{i+1} c_{i+1})\right) .$$

Remark, since ln is increasing, the dissipation D is nonnegative. Depending on the properties you are looking for, this is possible to define the *relative entropy* functional, with same dissipation term, and given by the expression

$$H(c|c^\rho) = \sum_{i\geq 1} c_i \left(\ln\left(\frac{c_i}{c_i^\rho}\right) - 1\right) + \sum_{i\geq 1} c_i^\rho ,$$

where c^ρ is the equilibrium candidate, with mass ρ, i.e. the components are given by Eq. (8) for which z is chosen such that $\sum_{i\geq 1} i Q_i z^i = \rho$. The second term in the right-hand side, ensuring non-negativity, is sometimes omitted. Hence, in the case initially $H(c^{\text{in}}|c^\rho) < +\infty$, we should have $D(c(t)) \to 0$ as $t \to \infty$ as we can see in

$$0 \leq H(c(t)|c^\rho) + \int_0^t D(c(s))\, ds \leq H(c^{\text{in}}|c^\rho) . \tag{10}$$

And remarking that $D = 0$ corresponds, see its definition, to $J_i = 0$ for all $i \geq 1$, we have a good reason to go ahead with the functional H. The hard work is to prove rigorous properties on the entropy and relative entropy, along the solutions. Again Ball et al. (1986) set the basic. The authors give many results, among others, continuity properties of the entropy functional (Proposition 4.5) and minimizing sequence properties (Theorem 4.4). Also, they proved the key ingredient that Eq. (10) holds for a large class of rate (Theorem 4.8). Finally, in their Theorem 4.7, they proved the so-called H-theorem (by analogy with the celebrating Boltzmann H-theorem), which is a rigorous justification of (9).

Theorem 2 (H-Theorem, Ball et al. 1986) *Assume* $z_s > 0$, $\liminf_{i\to\infty} Q_i^{1/i} > 0$, $a_i = O(i/\ln i)$ *and* $b_i = O(i/\ln i)$. *If* c *is a solution to the Becker–Döring equations* (4)–(5) *on* $[0, T)$, *for some* $T \in (0, +\infty]$, *with initial condition* $c^{\text{in}} \neq 0$ *belonging to* X^+, *then dissipation of entropy Eq.* (9) *holds almost every* $t \in [0, T)$.

Note that linear growth a_i, $b_i \sim i$ is not allowed. Fewer assumptions on the rate of fragmentation is possible, adapting the results obtained for general coagulation-fragmentation equation by Carr and da Costa (1994) and later by Cañizo (2007). Now we state the main asymptotic results. A very general result in the case $z_s = +\infty$ is available from Theorem 5.4 by Ball et al. (1986). But the more interesting case is $0 < z_s < +\infty$ for which a dichotomy occurs. This is treated for particular initial conditions and rates in Ball et al. (1986), and then extended to general initial

conditions by Ball and Carr (1988). Finally, it was refined by Slemrod (1989) for a class of rates allowing linear growth, see its Theorem 5.11, which we state below.

Theorem 3 (Convergence to Equilibrium, Slemrod 1989) *Let $c^{\text{in}} \in X^+$ with mass ρ and such that $H(c^{\text{in}}) < +\infty$. Assume $a_i = O(i)$, $b_i = O(i)$ and $\lim_{i\to+\infty} Q_i^{1/i} = 1/z_s$ exists ($z_s > 0$). Assume moreover there exists $z \in [0, z_s]$ such that $a_i z \leq b_i$ for sufficiently large i. Finally, let c be the unique solution to the Becker–Döring equations (4)–(5) on $[0, +\infty)$ with initial data c^{in}. We have:*

(a) *If $0 \leq \rho \leq \rho_s$, then* $\lim_{t\to+\infty} \sum_{i\geq 1} i|c_i(t) - c_i^\rho| = 0$.

(b) *If $\rho > \rho_s$, then, for every $i \geq 1$,* $\lim_{t\to+\infty} c_i(t) = c_i^{\rho_s}$.

In both case we recall that c^ρ is the equilibrium given by Eq. (8) with mass ρ.

Surprisingly, in point (b), while the solution has mass ρ for all times, as the time goes to infinity, it converges in a weak sense (component by component) to a solution having a strictly inferior mass. In this theory, the difference $\rho - \rho_s$ is interpreted as the formation of particles with *infinite* size and of different nature, phenomenon called *phase transition*. In Sect. 2.3.1 we will describe in more detail this phenomenon.

The proof consists first in proving that the $w(c^{\text{in}})-$limit set consist of equilibrium candidate $c^{\rho'}$ with mass ρ' less than $\min(\rho, \rho_s)$. This is achieved by compactness of the orbit, analyzing the time-translation, and by regularity of the c_1 which requires in particular $b_i = O(i)$ (see Theorem 3.2 in Ball et al. 1986), contrary to the known existence result stated above in Theorem 1. Then, the limit is selected thanks to the dissipation (10). A key ingredient is the continuity property of $c \mapsto H(c|c^\rho)$ which holds if and only if $\lim_{i\to+\infty} Q_i^{1/i}$ exists and $\rho = \rho_s$, see Proposition 4.5 by Ball et al. (1986). Note, the condition $a_i z \leq b_i$ comes from the Theorem 2 by Ball and Carr (1988), and re-used by Slemrod (1989), which ensures the tail of the c_i's decays sufficiently fast (fragmentation dominates). We point out that these two last conditions are needed to select the right equilibrium, while convergence to some equilibrium is "always" satisfied, see Theorem 5.10 in Slemrod (1989).

We finish by a comment on the case where $z_s = 0$, corresponding to a strong coagulation rate, relatively to the fragmentation. Carr and Dunwell (1999) proved under reasonable assumptions that for all $i \geq 1$, $c_i(t) \to 0$ as $t \to +\infty$.

2.2.2 Rate of Convergence

The natural question that arises after the convergence to equilibrium is the *rate* of convergence. When the H-Theorem 2 holds, with the relative-entropy for instance, we could hope that convergence holds in this sense. The best situation would be the dissipation bounded from below by the entropy itself, i.e. along the solutions: $D(c(t)) \geq CH(c(t)|c^\rho)$ for some constant $C > 0$. This leads immediately to an exponential decay of the entropy. Unfortunately this does not hold in all cases.

A recent proof for $a_i \sim i$ is given by Cañizo et al. (2015). Another way is to bound from below the dissipation by a non-negative function ψ depending on H, leading to

$$\frac{d}{dt}H(c(t)|c^\rho) \leq -\psi(H(c(t)|c^\rho))\,. \tag{11}$$

And the problem resumes to find sub-solutions to this ordinary differential equation. This method is named *entropy entropy-dissipation*, because dissipation is created by entropy itself. But this method does not in general lead to exponential decay of the entropy. The first result in this direction is due to Jabin and Niethammer (2003). Let us show their result.

Theorem 4 (Rate of Convergence, Jabin and Niethammer 2003) *Assume* $1 \leq a_i = O(i)$, $1 \leq b_i = O(i)$, $\lim_{i\to+\infty} Q_i^{1/i} = 1/z_s$ *exists* ($z_s > 0$) *and that* $a_i z_s \leq \min(b_i, b_{i+1})$ *for every* $i \geq 1$. *Suppose moreover that* $c^{\text{in}} \in X^+$ *with mass* $\rho < \rho_s$ *(sub-critical case), with* $H(c^{\text{in}}|c^\rho) < +\infty$ *and there exists* $\nu > 0$ *such that* $\sum_{i\geq 1} \exp(\nu i) c_i^{\text{in}} < +\infty$. *The solution* c *to the Becker–Döring equations* (4)–(5) *on* $[0, +\infty)$ *with initial data* c^{in} *satisfies, for some constant* k *depending on* c^{in} *and for all* $t \geq 0$

$$H(c(t)|c^\rho) \leq H(c^{\text{in}}|c^\rho)\exp(-kt^{1/3})\,.$$

This theorem is obtained thanks to the possibility to choose $\psi(H) = H/\ln(H)^2$ in Eq. (11). And the authors were able to go back from this estimate to the convergence in $\exp(-kt^{1/3})$ in the strong norm of X^+. Similar results are obtained by Cañizo et al. (2015) in various cases allowing fewer hypotheses. But these results still not provide satisfactory rate of decay, with pure exponential decay. A well-know theory is the stability of linear operator. If the linearized system is locally exponentially stable, we could hope that so is the full non-linear system, in a small neighborhood of the equilibrium. And we could imagine that this small neighborhood is an absorbing set, since we have Theorem 4. In fact these steps were followed by Cañizo and Lods (2013) to obtain their nice proof of the full exponential convergence stated below.

Theorem 5 (Exponential Stability, Cañizo and Lods 2013) *Under the hypothesis of Theorem 4 and in addition* $\lim_{i\to+\infty} a_{i+1}/a_i = z_s \cdot \lim_{i\to\infty} Q_{i+1}/Q_i = 1$. *The solution* c *to the Becker–Döring equations* (4)–(5) *on* $[0, +\infty)$ *with initial data* c^{in} *satisfies, for all* $t \geq 0$,

$$\|c(t) - c^\rho\| \leq A\exp(-\lambda t)\,,$$

for some constant $A > 0$.

The constant λ is completely calculable from the constant of the problem, important fact for applicability. We refer to their article (Cañizo and Lods 2013) for a very well-detailed introduction and presentation to the result. We mention that the linear Becker–Döring system (with constant monomer c_1) also exhibits exponential decay (Kreer 1993), and a quantitative comparison of the convergent rates will be of interest.

We finish this section pointing out that whether exponential convergence towards the steady state holds true for either the critical case or the super-critical case still remains open. We will detail in the Sect. 2.4 that metastability phenomena are present in the super-critical case.

2.3 Coarsening and Relation to Transport Equation

From the Becker–Döring equations (4)–(5), the reader familiar with numerical analysis may recognize that equations on c_i, for every $i \geq 2$, has the flavor of a discretization of a transport equation. To make the link more apparent, it is useful to write down the weak form of Eqs. (4)–(5), which is also a very useful tool for the study of the BD system it-self. Take $(\varphi_i)_{i\geq 2}$ a sufficiently regular sequence, we then obtain

$$\frac{d}{dt}\sum_{i\geq 2} c_i(t)\varphi_i = \varphi_2 J_1 + \sum_{i\geq 2}(\varphi_{i+1}-\varphi_i)\, J_i\,. \tag{12}$$

where we recall the fluxes J_i are defined by Eq. (3). Clearly, $(\varphi_{i+1}-\varphi_i)$ can be seen as a discrete "spatial" derivative. Moreover, assuming some "spatial continuity," it is tempting to rewrite J_i as $J_i \approx (a_i c_1 - b_i)\, c_i$. With such ansatz, the last equation (12) then motivates the introduction of the following continuous transport equation (in a weak form)

$$\frac{d}{dt}\int_0^\infty \varphi(x) f(t,x)dx = \varphi(0)N(t) + \int_0^\infty \varphi'(x) j(t,x) f(t,x)dx\,, \tag{13}$$

where the flux now reads

$$j(t,x) = a(x)u(t) - b(x)\,,$$

for some appropriate functions a and b, and a function u that plays the role of c_1. We will see later what should be N and what becomes the mass conservation stating u in the subsequent sections. Both are the main difficulties of the problem in linking the discrete Eq. (12) to the continuous Eq. (13).

As a matter of fact, they depend crucially on the scaling hypothesis (a small parameter which allows passing from discrete size i to continuous size x) and on the kinetic coefficients a and b. We note that Eq. (13) is the weak form of a nonlinear transport equation known as the Lifshitz–Slyozov (LS) equation, after the work by Lifshitz and Slyozov (1961),

$$\frac{\partial}{\partial t} f + \frac{\partial}{\partial x}\big(j(t,x) f(t,x)\big) = 0\,, \tag{14}$$

together with (if appropriate) the boundary condition at $x = 0$,

$$\lim_{x \to 0^+} j(t,x) f(t,x) = N(t)\,, \tag{15}$$

and an equation for u. Rigorous results making connection from the Becker–Döring Eqs. (4)–(5) to the Lifshitz–Slyozov Eq. (14) are of two kinds. First, in the works initiated by Laurençot and Mischler (2002) and Collet et al. (2002), and pursued in Deschamps et al. (2017), the authors proved that a suitable rescaling of the solution to BD equations (with the essential assumption of large excess of monomers c_1) converges to a solution of LS, on any finite time period and either in density or measure functional spaces. Second, in the works initiated by Penrose et al. (1978) and pursued by Penrose (1997), Niethammer (2003; 2004), the authors show that long-time behavior of super-critical solutions to BD equations are closed to the solution of LS.

2.3.1 Evolution of Large Clusters in the Super-Critical Case

We saw in Theorem 3, in the case $\rho > \rho_s$, that the solution behaves particularly, as infinitely large clusters are created as time goes to infinity. The idea by Penrose (1997) is to perform a time/space scaling to approach the cluster distribution, both in a very long time and for very large sizes, in order to explain the loss of mass $\rho - \rho_s > 0$ in the super-critical case. The formal arguments for coefficients given by Eq. (2) with $\gamma = \alpha = 1/3$ are derived in Penrose (1997), we refer also to the review by Slemrod (2000). We present here the rigorous result obtained by Niethammer (2003), for any coefficients given by Eq. (2) where the author proved that large clusters obey a variant of LS, named the Lifshitz–Slyozov–Wagner (LSW) equations, see below. For a review on LSW and Ostwald Ripening (out of the scope of this paper), see Niethammer et al. (2006) and Niethammer (2008).

We sketch the formal arguments following Penrose (1997) and Niethammer (2003). To consider the behavior of large clusters at large time, we introduce an ad hoc small parameter $0 < \varepsilon \ll 1$ such that $1/\varepsilon$ will be a measure of a typical large cluster. Within the particular choice of coefficients given by Eq. (2), it turns that a new time scale given by $\tau = \varepsilon^{1-\alpha+\gamma} t$, where α and γ are the exponents arising in the coefficients, is an appropriate time scaling to obtain a nontrivial dynamics. Indeed, we obtain by the BD equations (4)–(5) the reformulation

$$\frac{d}{d\tau} c_i = \frac{1}{\varepsilon^{1-\alpha+\gamma}} \left(J_{i-1} - J_i \right)\,,$$

and the fluxes J_i in Eq. (3) become

$$J_i = a_i \left(c_1 - z_s - \frac{q}{i^\gamma} \right) c_i - \left(b_{i+1} c_{i+1} - b_i c_i \right)\,,$$

for every $i \geq 1$. Since large clusters are formed as time goes to infinity, it is possible to consider the system after a (possibly long) time t_ε for which the relative entropy $H(c|c^{\rho_s})$ is small enough, namely of order ε^γ. This suggests that the small clusters, up to some cut-off i_ε are close to their equilibrium value, for $t \geq t_\varepsilon$ and $i \leq i_\varepsilon$, given by

$$c_i(t) = Q_i z_s^i (1 + o(1)) .$$

On the other hand, large cluster may be described by a continuous variable $x = \varepsilon i$ for $i \geq i_\varepsilon$. Thus, we define a density f (stepwise) according to the variable $x \geq \varepsilon i_\varepsilon$ by

$$f^\varepsilon(\tau, \varepsilon i) = \frac{1}{\varepsilon^2} c_i(\tau) .$$

Respectively, we let $u^\varepsilon(\tau) = (c_1(t) - z_s)/\varepsilon^\gamma$. This yields, after some manipulations,

$$\frac{\partial f^\varepsilon(\tau, x)}{\partial \tau} + \frac{j^\varepsilon(\tau, x-\varepsilon) f^\varepsilon(\tau, x-\varepsilon) - j^\varepsilon(t, x) f^\varepsilon(t, x)}{\varepsilon} = o(1) ,$$

with $j^\varepsilon(\tau, x) = x^\alpha \left(u^\varepsilon(\tau) - \frac{q}{x^\gamma}\right)$. Formal arguments lead, as $\varepsilon \to 0$, to a solution f of Eq. (14). In turn, the mass conservation (6) becomes

$$\rho = \sum_{i=1}^{\infty} i c_i(\tau) = \sum_{i=1}^{i_\varepsilon} i c_i(\tau) + \sum_{i=i_\varepsilon}^{\infty} i c_i = \rho_s + \int_0^\infty x f(\tau, x)\, dx + o(1) ,$$

At the limit, we obtain $\int_0^\infty x f(\tau, x)\, dx = \rho - \rho_s$ which measures large clusters formation. Such condition, complemented with the LS equation (14), allows determining u by the following expression

$$u(\tau) = \frac{q \int_0^\infty x^{\alpha-\gamma} f(\tau, x)\, , dx}{\int_0^\infty x^\alpha f(\tau, x)\, dx} .$$

We now state the result obtained by Niethammer (2003).

Theorem 6 (Lifschitz–Slyozov–Wagner Limit, Niethammer 2003) *Assume kinetic coefficient are given by Eq. (2), that the initial condition $c^\varepsilon(0)$ satisfies $H(c^\varepsilon(0)|c^{\rho_s}) = \varepsilon^\gamma$ and that $\sum_{i \geq M/\varepsilon} i c_i^\varepsilon(0) \to 0$ as M goes to infinity uniformly in $\varepsilon > 0$.*

There is a subsequence $\{\varepsilon_n\}$ converging to 0, a measure-valued function $t \mapsto \nu_t$ solution of LS equation (14) in $\mathcal{D}'(\mathbb{R}_+ \times (0, +\infty))$ such that

$$\int_0^\infty \varphi(x) f^\varepsilon(\tau, x)\, dx \to \int_0^\infty \varphi(x) \nu_t(dx) ,$$

locally uniformly in $t \in \mathbb{R}^+$, *for all* $\varphi \in C_0^0(0,\infty)$ *and for all* $t \geq 0$, *and*

$$\int_0^\infty x\nu_t(dx) = \rho - \rho_s\,.$$

We also mention the case of vanishing small excess of density, $\rho - \rho_s \to 0$ as $\varepsilon \to 0$, by Niethammer (2004), where the authors recovered the LS equation, in a similar framework.

2.3.2 Rescaled Solution of BD for Large Monomer Density

Another point of view is to consider fast reaction rates $a_i c_1 c_i \sim b_{i+1} c_{i+1}$ of order $1/\varepsilon$, where $0 < \varepsilon \ll 1$, together with a large excess of monomers. Namely, the characteristic number of free particles c_1 is two orders of magnitude greater than the characteristic number of clusters with size $i \geq 2$. Following Collet et al. (2002), alternatively Deschamps et al. (2017), this leads to a rescaled version of the BD equations (4)–(5) given, for $\varepsilon > 0$, by

$$\frac{d}{dt}u^\varepsilon = -\varepsilon J_1^\varepsilon - \varepsilon \sum_{i\geq 1} J_i^\varepsilon\,, \tag{16}$$

$$\frac{d}{dt}c_i^\varepsilon = \frac{1}{\varepsilon}\Big[J_{i-1}^\varepsilon - J_i^\varepsilon\Big]\,, \tag{17}$$

for every $i \geq 1$, where u^ε is the dimensionless version of c_1 (not to be confused with the previous section) and the scaled fluxes are

$$J_1^\varepsilon = \alpha^\varepsilon (u^\varepsilon)^2 - b_2^\varepsilon c_2^\varepsilon\,, \qquad J_i^\varepsilon = a_i^\varepsilon u^\varepsilon c_i^\varepsilon - b_{i+1}^\varepsilon c_{i+1}^\varepsilon\,,$$

for every $i \geq 1$. Theorem 1 provides existence and uniqueness of solution at fixed $\varepsilon > 0$. Collet et al. (2002) constructed a sequence of "density" approximations in the Lebesgue space $L^1(\mathbb{R}_+)$ by, for all $t \geq 0$ and $x \geq 0$

$$f^\varepsilon(t,x) = \sum_{i\geq 2} c_i^\varepsilon(t)\mathbf{1}_{\Lambda_i^\varepsilon}(x)\,,$$

where $\Lambda_i^\varepsilon = [(i-1/2)\varepsilon, (i+1/2)\varepsilon)$ for each $i \geq 2$. Note the first cluster is excluded from the density, it is like assuming a solute with density f^ε belonging to the solvent u^ε (in large excess). Then, macroscopic aggregation and fragmentation rates are constructed as functions on $\mathbb{R}_+$ (similarly to f^ε), for each $\varepsilon > 0$ and $x \geq 0$,

$$a^\varepsilon(x) = \sum_{i\geq 2} a_i^\varepsilon \mathbf{1}_{\Lambda_i^\varepsilon}(x)\,, \qquad b^\varepsilon(x) = \sum_{i\geq 2} b_i^\varepsilon \mathbf{1}_{\Lambda_i^\varepsilon}(x)\,.$$

This scaling supposes the first coagulation rate α^ε is faster (order $1/\varepsilon^2$) than the other rates a_i^ε for $i \geq 2$, which justifies the use of another notation α^ε and a special treatment outside the function a^ε. Theoretical justifications can be found in Collet et al. (2002). Finally, the balance of mass reads in this case, for all $t \geq 0$

$$u^\varepsilon(t) + \int_0^\infty x f^\varepsilon(t,x)\,dx = \rho^\varepsilon\,, \tag{18}$$

for some $\rho^\varepsilon > 0$. The value of ρ^ε is entirely determined by the initial condition at time $t = 0$.

Again we deal with the limit $\varepsilon \to 0$, and we hope the limit of f^ε satisfies in some sense the LS equation (14). Let us introduce few hypotheses for the limit theorem, namely we assume, there exists a constant $K > 0$, independent on $\varepsilon > 0$, such that, for all $x \geq 0$,

$$a^\varepsilon(x) + b^\varepsilon(x) \leq K(1+x)\,. \tag{19}$$

Also, we assume there exists a measure μ^{in} on $\mathbb{R}_+$ such that

$$\lim_{\varepsilon\to 0} \int_0^\infty \varphi(x) f^\varepsilon(0,x)\,dx = \int_0^\infty \varphi(x)\mu^{\text{in}}(dx)\,, \tag{20}$$

for all $\varphi \in \mathcal{C}_0((0,+\infty))$ and

$$\lim_{R\to+\infty} \sup_{\varepsilon>0} \int_R^\infty x f^\varepsilon(0,x)\,dx = 0\,. \tag{21}$$

This estimate on the tail of the initial distribution is a classical argument which increases the compactness and will allow then to pass to the limit in the balance of mass (18).

Finally, we resume in the following the results obtained by Collet et al. (2002) in their Theorem 2.3, by Laurençot and Mischler (2002) in Theorem 2.2 for a different framework, and also modified by Deschamps et al. (2017), in Lemma 5.

Theorem 7 (Lifschitz–Slyozov Limit, Collet et al. 2002, Laurençot and Mischler 2002) *Assume that α^ε is uniformly bounded, and that a^ε and b^ε satisfy Eq. (19). Suppose moreover that there exists $\rho \geq 0$ and two non-negative real functions a and b defined on $\mathbb{R}_+^*$ such that, when ε converges to 0, ρ^ε converges to ρ, a^ε and b^ε converge locally uniformly on $\mathbb{R}_+^*$ toward, respectively, a and b.*

If the family $\{f^\varepsilon(0,\cdot)\}$ satisfies Eqs. (20) and (21), then from all sequences $\{\varepsilon_n\}$ converging to 0 we can extract a subsequence still denoted $\{\varepsilon_n\}$ such that

$$\lim_{n\to\infty} \int_0^\infty \varphi(x) f^{\varepsilon_n}(t,x)\,dx = \int_0^\infty \varphi(x)\mu(t,dx)\,, \tag{22}$$

locally uniformly in $t \in \mathbb{R}_+$, *and for all* $\varphi \in \mathcal{C}_0(0, +\infty)$, *where* $\mu := (\mu(t,\cdot))_{t\geq 0}$ *is a measure-valued function satisfying the LS equation* (14) *in* $\mathcal{D}'(\mathbb{R}_+ \times (0, +\infty))$ *where* $u \in \mathcal{C}(\mathbb{R}_+)$ *is non-negative and satisfies, for all* $t \geq 0$,

$$u(t) + \int_0^\infty x\mu(t, dx) = \rho\,. \tag{23}$$

The proof relies, mainly, on moment estimates and equicontinuity arguments. This theorem does not conclude on the full convergence of the family as $\varepsilon \to 0$. To that it requires a uniqueness argument of the limit problem Eq. (14) in measure with the balance of mass (23). Looking Eq. (14) against functions in $\mathcal{D}(\mathbb{R}_+ \times (0, +\infty))$ allows uniqueness with the necessary condition that the flux $j(t, x)$ points outward the domain at $x = 0$ (for instance, if $a(0)\rho - b(0) < 0$). We refer to the works by Niethammer and Pego (2000), Collet and Goudon (2000) and by Laurençot (2001) for the well-posedness theory on the Lifshitz–Slyozov equation. Also, we mention that the convergence in Eq. (22) has also been shown to hold in a functional density space, in $L^1(xdx)$, by Laurençot (2002).

We are now concerned with the case the flux $j(t, x)$ points inward the domain at $x = 0$, for instance if $a(0)u(0) - b(0) > 0$, or more generally if the characteristics, backward solution of

$$\frac{d}{dt}x = j(t, x)$$

goes back to $x = 0$ in finite time. In this case, it is hopeless to obtain a well-defined limit to the LS equation (14) without a boundary condition, of type (15). A rigorous identification of the boundary condition has been performed by Deschamps et al. (2017). It was obtained through the limit of the rescaled BD equations (16)–(17) in the spirit of Theorem 7. More precisely, we assumed, $a(x) \sim_0 \bar{a}x^{r_a}$ and $b(x) \sim_0 \bar{b}x^{r_b}$ with $r_a \leq r_b$ and $r_a < 1$. These assumptions allow a fine control of the pointwise value of the solution at $x = 0$ to obtain the boundary value. The limit obtained is a measure-valued solution to LS on $[0, T]$, identifiable if $\sup_{t\in[0,T]} u(t) > \lim_{x\to 0} b(x)/a(x)$ which corresponds to time interval on which characteristic goes back to $x = 0$. Let us present an informal version of a result we obtained.

Theorem 8 (Boundary Value, Deschamps et al. 2017) *A "good" boundary condition at* $x = 0$ *for the Lifschitz–Slyozov equation, when* $a(x) = \bar{a}x^{r_a}$ *and* $b(x) = \bar{b}x^{r_b}$ *with* $r_a < 1$ *and* $r_a \leq r_b$, *is*

$$\lim_{x\to 0^+} j(t, x)f(t, x) = \begin{cases} \alpha u(t)^2\,, & \text{if } r_a < r_b\,,\ u(t) > 0\,; \\ \frac{\alpha}{\bar{a}}u\left(\bar{a}u - \bar{b}\right)\,, & \text{if } r_a = r_b\,,\ u(t) > \bar{b}/\bar{a}\,, \end{cases}$$

where α is the limit of α^ε as ε goes to 0. In both cases, this also reads

$$\lim_{x\to 0^+} x^{r_a} f(t,x) = \frac{\alpha}{\overline{a}} u(t)\,.$$

Note the conditions on r_a, r_b, and u are well related to incoming characteristic. Theorems 1 and 2 by Deschamps et al. (2017) also assumed a technical growth condition (in ε) on the "relatively small" sizes, through the condition

$$\sup_{\varepsilon>0} \sum_{i\geq 2} \varepsilon^{r_a} c_i^{\text{in},\,\varepsilon} e^{-iz} < +\infty\,,$$

for all $z \in (0,1)$. This is the key estimates which is proved to propagate in time (see Proposition 2). This allows a quasi-steady-state limit of the small cluster concentrations, that behave as fast variables in Eq. (17). Note in the case of exact power law, Deschamps et al. (2017) also proved with extra reasonable assumptions on initial conditions, that the limit measure solution has a density with respect to $x^{r_a}dx$. Finally, other scalings of the first fragmentation rate are investigated in Deschamps et al. (2017). Also, these results do not provide a complete answer. Indeed, uniqueness for the inward case is not achieved and we are not aware if u can cross the threshold $\lim_{x\to 0} b(x)/a(x)$.

Remark 1 Second-order approximations (Fokker–Planck like) of BD equations are still under intense active research, and a full satisfactory answer is still an open problem, see proposed equations by Velázquez (1998; 2000), Hariz and Collet (1999), Collet et al. (2002), Collet (2004), Conlon et al. (2016). Arbitrary higher order terms are formally derived by Niethammer (2003).

2.4 Time-Dependent Properties, Metastability and Classical Nucleation Theory

The following properties are of the most important ones in application of the BD equations to phase transition. Yet, as for the convergence rate to equilibrium, coarsening and evolution of large sized clusters, available results are still incomplete. The main result we are aware of on metastability for BD equations (4)–(5) is given by Penrose (1989). The ideas of classical nucleation theory goes back to Becker and Döring (1935), and is built on the remark that there exist steady-state solutions of Eq. (5) (with c_1 constant) with non-zero steady-state flux, which can be arbitrarily small in some sense. This very small steady-state flux is interpreted as the rate of formation of larger and larger cluster, leading to a phase transition phenomena in long time. The term metastability in such theory refers to the fact that the rate is arbitrary small. Penrose (1989) goes much beyond by extending this notion of metastability to a time-dependent phenomenon (instead of a steady-state one). Indeed, he could exhibit a solution of the full system (4)–(5) that enters a state that

lived for exponentially long time, yet can be distinguished from the equilibrium state. This solution is a super-critical solution, with $\rho > \rho_s$, and is required to have a well prepared initial condition. This solution is also related in some sense to an extremely small common flux value. It remains an important open question to know whether the metastable state can be reached from a larger class of initial data.

Penrose (1989) considered technical conditions on coefficients which are essentially satisfied by the ones given by Eq. (2). The crucial initial condition is then constructed as follows. For any $z > z_s$, let $f_i(z)$ be the unique solution of

$$a_{i-1} z f_{i-1}(z) - (b_i + a_i z) f_i(z) + b_{i+1} f_{i+1}(z) = 0\,, \quad i \geq 2\,,$$

with end conditions $f_1(z) = z$ and $\sup_i f_i(z) < \infty$. Actually, f_i can be solved explicitly by (for $z > z_s$ the reader can check that the infinite series are convergent)

$$f_i(z) = J(z) Q_i z^i \sum_{r=i}^{\infty} \frac{1}{a_r Q_r z^{r+1}}\,, \quad J(z) := \left[\sum_{r=1}^{\infty} \frac{1}{a_r Q_r z^{r+1}} \right]^{-1}.$$

Let i^* be the critical cluster size defined as the (unique) size that minimizes the quantity $a_i Q_i z^i$. The metastable state exhibited by Penrose (1989) has to be understood in the limit of small excess of density, $z \searrow z_s$. The following terminology is used

- $g(z)$ is exponentially small if for each $m > 0$, $g(z) = O((z - z_s)^m)$.
- $g(z)$ is at most algebraically large if for some $m > 0$, $g(z) = O((z - z_s)^{-m})$.

The main theorem by Penrose (1989) reads

Theorem 9 (Metastability, Penrose 1989) *Let c be the solution of the BD Eqs. (4)–(5) with initial condition*

$$c_i(0) = \begin{cases} f_i(z)\,, & \text{if} \quad i \leq i^*\,, \\ Q_i z_s^i\,, & \text{if} \quad i > i^*\,. \end{cases}$$

Then c has an exponentially long lifetime as $z \searrow z_s$, in the sense that for each fixed i (note that $i^* \to \infty$):

- *if t is at most algebraically large, then $c_i(t) - c_i(0)$ is exponentially small*
- *$\lim_{t\to\infty}[c_i(t) - c_i(0)]$ is not exponentially small*

Thus, cluster with size $i \ll i^*$ remain exponentially close to their initial values, until an exponentially long time has elapsed. But eventually they do change. Note that the initial values for the small clusters, $f_i(z)$, correspond to the steady-state values of the classical nucleation theory, for which $J_{i-1}(0) = J_i(0)$ for all $2 \leq i < i^*$, and the common flux value is $J(z)$, which is also exponentially small as $z \searrow z_s$. We refer the reader to Penrose (1989, Theorems 1 and 2) for orders of magnitude of i^*, $J(z)$ and quantification of the (small) growth rate of large clusters of size greater than i^*.

Remark 2 The numerical illustration of the metastability is a problem per se, we refer the reader to the two nice papers by Carr et al. (1995) and by Duncan and Soheili (2001), where numerical schemes are derived and are shown to consistently represent the metastable states. The reader may also look at the Sect. 3.4 where numerical simulations of the stochastic Becker–Döring are shown. Finally, let us mention that analogous metastability properties have been investigated in the classical linear version of BD by Penrose (1989) and Kreer (1993), in a finite-dimensional truncated version by Dunwell (1997) and Duncan and Dunwell (2002), and in a thermodynamically consistent version of the BD system by Ssemaganda and Warnecke (2013).

3 Stochastic Becker–Döring Model

Due to space considerations, we will not detail historical facts on the study of stochastic coagulation-fragmentation models. Let us just mention that the first study of a stochastic coagulation models is widely attributed to Marcus (1968) and Lushnikov (1978) which give the name to the Marcus–Lushnikov process, stochastic analog of the pure coagulation Smoluchowski's equations. Up to our knowledge, Whittle (1965) and Kelly (1979) are pioneers in the study of more general stochastic coagulation-fragmentation models (including the Becker–Döring model). See Aldous (1999) and the discussion by Freiman and Granovsky (2005) for more details.

3.1 Definition and State-Space

A stochastic version of the Becker–Döring model may be defined as a continuous time Markov chain analog of the set of ordinary differential equations (4)–(5), for which transitions are given by the same set of kinetic reactions (1), but modeling *discrete numbers* of clusters instead of continuous concentrations. Precisely, given a positive integer M, we define the state space

$$X_M := \left\{ C = (C_i)_{i\geq 1} \in \mathbb{N}^{\mathbb{N}} \,:\, \sum_{i=1}^{M} iC_i = M \right\} .$$

On X_M, we introduced the following operators defined by, for any configuration C on X_M,

$$R_1^+ C = (C_1 - 2\,, C_2 + 1\,, \ldots\,, C_i\,, \ldots)$$
$$R_2^- C = (C_1 + 2\,, C_2 - 1\,, \ldots\,, C_i,, \ldots)$$

and, for any $i \geq 2$,

$$R_i^+ C = (C_1 - 1\,, C_2\,, \ldots\,, C_i - 1\,, C_{i+1} + 1\,, \ldots)$$
$$R_{i+1}^- C = (C_1 + 1\,, C_2\,, \ldots\,, C_i + 1\,, C_{i+1} - 1\,, \ldots)$$

Given non-negative kinetic rates $(a_i)_{i\geq 1}$, $(b_i)_{i\geq 2}$, the stochastic Becker–Döring model (SBD) is defined as the continous time Markov chain on X_M with transition rates

$$\begin{cases} q(C, R_1^+ C) = a_1 C_1 (C_1 - 1)\,, \\ q(C, R_i^+ C) = a_i C_1 C_i\,, & i \geq 2\,, \\ q(C, R_i^- C) = b_i C_i\,, & i \geq 2\,. \end{cases}$$

Given an initial configuration $C^{\text{in}} \in X_M$ (deterministic or random), the configuration $C(t)$ defined by the SBD may alternatively be represented as the solution of the following system of stochastic equations

$$\begin{cases} C_1(t) = C_1^{\text{in}} - 2J_1(t) - \displaystyle\sum_{i\geq 2} J_i(t)\,, \\ C_i(t) = C_i^{\text{in}} + J_{i-1}(t) - J_i(t)\,, \quad i \geq 2\,, \end{cases} \tag{24}$$

with

$$J_i(t) = Y_i^+\Big(\int_0^t a_i C_1(s)(C_i(s) - \delta_{1,i})ds\Big) - Y_{i+1}^-\Big(\int_0^t b_{i+1} C_{i+1}(s)ds\Big)\,, \quad i \geq 1\,,$$

where $\delta_{1,i} = 1$ if $i = 1$ and $\delta_{1,i} = 0$ if $i > 1$ and Y_i^+, Y_{i+1}^- for $i \geq 1$ are independent standard Poisson processes. Analogy between Eq. (24) and Eqs. (4)–(5) is clear. The number of clusters of size $i \geq 2$ evolves according to the differences between two (stochastic) cumulative counts J_{i-1} and J_i. Finally, we may also identify the SBD with the help of its infinitesimal generator L_M, defined by, for any bounded functions f on X_M,

$$L_M f(C) = \sum_{i=1}^{M-1} \big[f(R_i^+ C) - f(C)\big]\, a_i C_1 (C_i - \delta_{1,i}) + \big[f(R_{i+1}^- C) - f(C)\big]\, b_{i+1} C_{i+1}\,.$$

Thanks to the Markov processes theory, we deduce in particular that, for any bounded functions f on X_M,

$$f(C(t)) - f(C^{\text{in}}) - \int_0^t L_M f(C(s))ds$$

is a centered martingale, and, taking $f_C(C') = \mathbf{1}_{\{C'=C\}}$, we deduce the following Backward Kolmogorov equation on the probability $P(t,\cdot)$ on X_M (Master equation)

$$\frac{d}{dt}P(t;C) = \sum_{i=1}^{M-1} a_i(C_1+1)(C_i+1+\delta_{1,i})P(t;R_{i+1}^{-}C) - a_iC_1(C_i-\delta_{1,i})P(t;C)$$
$$+ \sum_{i=2}^{M} b_i(C_i+1)P(t;R_{i-1}^{+}C) - b_iC_iP(t;C)\,. \tag{25}$$

Although the well-posedness of the SBD model is of course standard (as a pure-jump Markov process on a finite state-space), a first nontrivial question arises with respect to the precise description of the state space, and in particular to its cardinality. In fact, the state space X_M is given by all possible partitions of the integer M, a well-known problem in combinatorics. In particular, one can show the recurrence formula and the asymptotic as $M \to \infty$, Flajolet and Sedgewick (2009, Chap. I.3)[1]

$$M \mid X_M \mid = \sum_{i=1}^{M} \sigma(i) \mid X_{M-i} \mid, \quad \mid X_M \mid \propto \frac{1}{4M\sqrt{3}} \exp\left(\pi\sqrt{\frac{2M}{3}}\right),$$

where $\sigma(i)$ is the sum of the divisors of i (e.g., $\sigma(6) = 1+2+3+6 = 12$).

Remark 3 We mention that some terminology in the literature may be confusing. Indeed, some authors (see Bhakta and Ruckenstein (1995)) have named the deterministic Becker–Döring system (4)–(5) a stochastic version of the Lifshitz–Slyozov(–Wagner) equations. Such terminology seems to be motivated by the fact that the size of clusters is modeled as discrete variable in Eqs. (4)–(5), and that such system has the "flavor" of a master equation for a random walk in $\mathbb{N}^+$.

Remark 4 As for the BD system (4)–(5), some variants have been considered for the SBD. Let us mention, for instance, the constant monomer system studied by Yvinec et al. (2016) (which leads to a Poissonian equilibrium distribution), the exchange-driven growth model studied by Ben-Naim and Krapivsky (2003) (where clusters exchange monomer in one single step), some reduced version for specific kinetic rates ($a_i = i, b_i = 0$) see Eugene et al. (2016) and Doumic et al. (2016) or for fixed number of clusters (only one or two clusters can be present) Yvinec et al. (2016), Penrose (2008) and Rotstein (2015). Of course, the SBD can be seen as a particular case of more general coagulation-fragmentation processes see (Bertoin 2006). However, due to its specificity, it seems that some of the results available on general coagulation-fragmentation processes are not straightforwardly applicable (or do not bring interesting conclusions). Finally, although out of the scope of this survey, let us mention the interesting links between the (stochastic) BD system with

[1] R. Yvinec thanks Bence Melykuti for pointing out this fact.

some lattice models (Chau et al. 2015; Dehghanpour and Schonmann 1997; den Hollander et al. 2000; Bovier et al. 2010; Ercolani et al. 2014), in particular for nucleation and phase transition.

3.2 Long-Time Behavior

Although Eq. (25) is linear with respect to $P(t,\cdot)$, the size of the state space being exponentially large as $M \to \infty$, it is illusory to obtain a full exact solution of Eq. (25). Yet, perhaps surprisingly, the stationary solution of Eq. (25) has a relatively simple form, namely a product-form (see Anderson et al. 2010). Indeed, the (unique) stationary probability Π on X_M of Eq. (25) is given by Kelly (1979, Theorem 8.1)

$$\Pi(C) = B_M \prod_{i=1}^{M} \frac{(Q_i)^{C_i}}{C_i!}\,, \tag{26}$$

where B_M is a normalizing constant and Q_i is defined by Eq. (8). One may verify simply that the following detailed balance condition holds (Kelly 1979, Theorem 1.2)

$$\Pi(C)q(C, R_i^+ C) = \Pi(R_i^+ C)q(R_i^+ C, C)$$

Note also that, for all $z > 0$, with $B_z := B_M/z^M$, the expression (26) may be rewritten $\Pi(C) = B_z \prod_{i=1}^{M} \frac{(Q_i z^i)^{C_i}}{C_i!}$, which has a clearer analogy with the deterministic equilibrium of the BD equation. Finally, the distribution Π has the following probabilistic meaning: let Z_i, $i = 1,\dots,M$, be independent Poisson random variables with respective means Q_i, then it is easily seen that, for all $C \in X_M$,

$$\Pi(C) = \mathbf{P}\left\{ Z_1 = C_1, \dots, Z_M = C_M \,\middle|\, \sum_{i=1}^{M} iZ_i = M \right\}\,.$$

For the stationary distribution Π, the expected number of clusters of size i is

$$\mathbf{E}_\Pi C_i = Q_i B_M / B_{M-i}\,,$$

and the probability that a randomly chosen particle lies in a cluster of size i is iQ_iB_M/MB_{M-i}, from which we deduce that the normalizing constant B_M satisfies the recursive formula (with $B_0 = 1$)

$$MB_M^{-1} = \sum_{i=1}^{M} iQ_i B_{M-i}^{-1}\,. \tag{27}$$

Moreover, B_M^{-1} is the coefficient of z^M in the power series expansion of

$$G(z) = \exp\left(\sum_i Q_i z^i\right)$$

Remark 5 In some examples, the recursive formula (27) may be solved exactly. For instance, if $a_i = ai$, $b_i = bi$, then the equilibrium probability is given by the closed-form formula

$$\Pi(C) = \binom{b/a + M - 1}{M}^{-1} \prod_{i=1}^{M} \frac{1}{C_i!} \left(\frac{b}{ai}\right)^{C_i},$$

Besides the analytical form of the equilibrium distribution Π, it is a natural question to ask what is its limiting behavior as $M \to \infty$. Under the assumption that

$$\lim_{i\to\infty} \frac{a_i}{b_{i+1}} = z_s > 0\,, \tag{28}$$

(which is slightly stronger than the hypothesis on $Q_i^{1/i}$ used in Theorems 2 and 3), one can show (see Freiman and Granovsky 2002; Bell and Burris 2003) that G has also for radius of convergence z_s, and in such case, the expected number of clusters of size i has a limit as $M \to \infty$, given by

$$\lim_{M\to\infty} \mathbf{E}_\Pi C_i = Q_i z_s^i\,, \tag{29}$$

Other functionals of the stationary distribution Π have been derived by Durrett et al. (1999). In particular, let us mention that the variance of C_i, under Π and with hypothesis (28), satisfies the same asymptotic relation (29), and that C_i, C_j, $i \neq j$, becomes asymptotically uncorrelated as $M \to \infty$. It is also interesting to note the link of the limit (29) with the supersaturation case in the deterministic BD theory, see Theorem 3. Study of the limit shape of the stationary distribution Π (and quantities like the size of the largest or lowest component) is a well-known problem in statistical physics or in combinatorics (study of random integer partitions and Young diagrams) and goes back to Khinchin's probabilistic method (Khinchin 1960). Detailed description of such field is out of the scope of this survey, and we refer the reader to Erlihson and Granovsky (2008), Freiman and Granovsky (2005), Granovsky (2013), Han et al. (2008) and Ercolani et al. (2014) for recent results.

In contrast to the deterministic theory, we are not aware of any work quantifying the speed of convergence toward the equilibrium distribution (26) (which has to be exponential). In particular, it would be interesting to study how this rate behaves as $M \to \infty$.

Remark 6 Strong binding limit for constant coefficients has been considered by D'Orsogna et al. (2012) (linked to the almost pure-coagulation deterministic

dynamics in King and Wattis (2002)) and illustrates how mass incommensurability arises for finite mass M, when a fixed maximal cluster size $N < M$ is further imposed.

3.3 Large Number and Relation to Deterministic Becker–Döring

A first natural question when comparing the SBD and the BD system is that can we recover the deterministic equations in the limit $M \to +\infty$? The main tool to answer such question is the tightness of stochastic processes, which provides an appropriate compactness property for a sequence of rescaled solutions of the SBD. As a particular case, Jeon (1998) has considered the sequence of stochastic processes $\{C^n(t)\}$ in $X_n^+ := \left\{\frac{1}{n}C : C \in \mathbb{N}^{\mathbb{N}}, \sum_{i\geq 1} iC_i = n\right\} \subset X^+ \subset l^2$, defined by the generator

$$L^n f(C) = n \sum_{i=1}^{n} \left[f(R_{i,n}^+ C) - f(C)\right] a_i C_1 (C_i - \delta_{1,i}) + \left[f(R_{i+1,n}^- C) - f(C)\right] b_{i+1} C_{i+1}\,, \tag{30}$$

where, for all $i \geq 1$,

$$\begin{aligned} R_{i,n}^+ C &= (C_1 - 1/n\,, C_2\,, \ldots\,, C_i - 1/n\,, C_{i+1} + 1/n\,, \ldots) \\ R_{i+1,n}^- C &= (C_1 + 1/n\,, C_2\,, \ldots\,, C_i + 1/n\,, C_{i+1} - 1/n\,, \ldots) \end{aligned}$$

Under such classical scaling (which satisfies the system size expansion), one can prove

Theorem 10 (Law of Large Numbers, Jeon 1998) *If a_i, b_i are such that*

$$\sup_{C\in X^+:\sum iC_i\leq 1} \sum_{i\geq 1} a_i C_i < \infty\,, \qquad \sup_{C\in X^+:\sum iC_i\leq 1} \sum_{i\geq 1} b_i C_i < \infty\,, \tag{31}$$

then the laws of the stochastic process $\{C^n(t)\}$ defined by Eq. (30) form a tight sequence as a càdlàg process in l^2.

Note that hypothesis (31) is trivially satisfied for sublinear function of i. Also, it is clear that any weak limit of $\{C^n(t)\}$ is a solution of the BD system (4)–(5), which is an alternative proof of existence of solution of the BD system. Finally, convergence of the whole sequence may be obtained with the uniqueness result stated in Theorem 1.

We are not aware of any rigorous derivation of a second-order approximation of such limit, which should reasonably be a Langevin stochastic differential equation version of the BD system (4)–(5).

3.4 Time-Dependent Properties, Metastability and Stochastic Nucleation Theory

Up to our knowledge, the early work by (Schweitzler et al. 1988) paves the way to study fluctuations of the time-dependent cluster distributions and first passage time in stochastic finite system nucleation models. Using physical arguments, they investigated reaction rates of the form $a_i \approx i^{2/3}$ and $b_i \approx i^{2/3} y_0 e^{qi^{-1/3}}$, which are asymptotically similar to Eq. (2) (which $\alpha = 2/3$, $\gamma = 1/3$). One can notice that for such coefficients, a (time-dependent) critical cluster size $i_c(t)$ exists, defined by

$$a_i C_1(t) - b_i < 0\,, \forall i < i_c\,, \quad a_i C_1(t) - b_i \geq 0\,, \forall i \geq i_c\,.$$

This observation has led (Schweitzler et al. 1988) to analyze the SBD with the Ostwald ripening theory in mind. Specifically, with the help of numerical simulations, and heuristically derived moment closure approximation of the master equation (25) governing the clusters' distribution evolution (which resembles second-order approximation of the deterministic BD system, see Remark 1), the authors put in evidence the existence of a (stochastic) metastable state which is reached before the equilibrium distribution. Indeed, starting from an initial pure-monomer condition, one can observe a rapid transient that leads to a relatively small cluster distribution (with support contained among the size below the critical size), which has a long-lived state. Only after a first critical cluster is formed, the cluster size distribution is bimodal, given by a mixture of undercritical and overcritical clusters, until a single large cluster emerges from a competition between overcritical clusters, and its further growth is at the expense of the other clusters which now shrink. We have reproduced similar numerical simulations, with kinetic coefficients given by Eq. (2), in Figs. 1 and 2.

A key event in exiting the metastable state is thus the formation of an overcritical cluster. Such event may be analyzed with the help of the first passage time theory. It is important to note that, in agreement with classical metastability theory, the authors of this previous work noticed that the first time needed to form an overcritical cluster was subjected to large fluctuations. We are not aware of any theoretical work on the metastability for the SBD system, but we may mention that several groups have recently investigated numerically the behavior of first passage time (or related quantities) in the SBD system (or related models) Bhatt and Ford (2003), Johansson (2016), Penrose (2008), Yvinec et al. (2012; 2016). In particular, it is tempting to use first passage time theory to define a stochastic analog of the so-called nucleation rate in the classical nucleation theory (see Sect. 2.4). However, we notice that the analytical form of such nucleation rate is unclear. In particular, what should be the quasi-stationary distribution, stochastic analogous to the metastable state derived in Sect. 2.4?

Finally, let us mention the link with the study of the stochastic gelation time in Smoluchowsky's coagulation model, which has recently been the subject of active research. Let us define, for $\alpha \leq 1$,

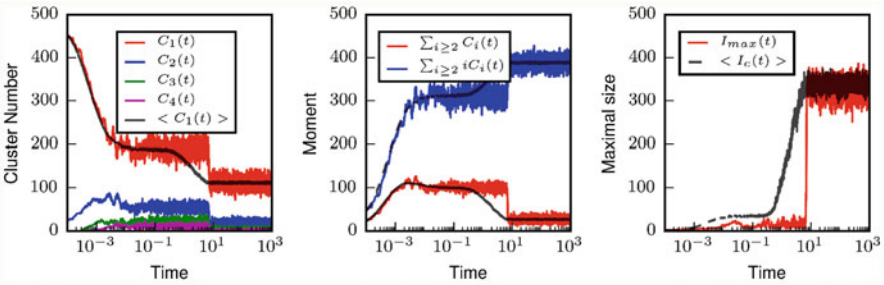


Fig. 1 Time trajectories of the SBD system (24) with kinetic coefficient given by (2), with $M = 500$, $\alpha = 2/3$, $\gamma = 1/3$, $z_s = 500/11$, $q = 10/11$. On the *left*, we plot a stochastic realization of the number of Monomers, Dimers, Tri-mers, and 4-mers, together with the sampled average over 100 realizations for the number of Monomers. On the *middle*, we plot the total mass in clusters and their numbers, and on the *right*, the maximal cluster size

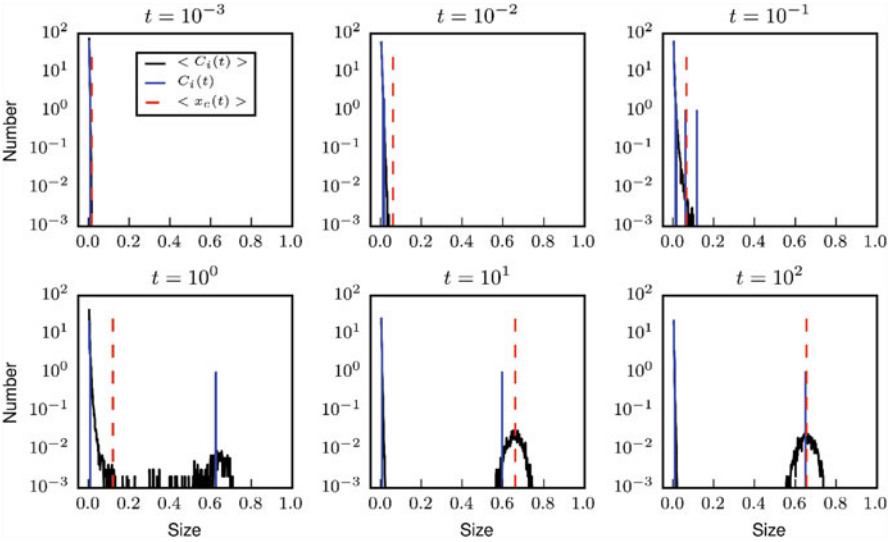


Fig. 2 Cluster size distribution at distinct times t, corresponding to Fig. 1. The sizes are rescaled by M, that is $x = 2/M, \ldots, 1$. In *blue* we plot the distribution of a stochastic realization, in *black* we represent the sampled averaged distribution over 1000 realizations, and in *red* we plot the (rescaled) critical size $x_c = I_c/M$

$$\tau_n^\alpha = \inf\{t > 0 \,:\, C_k^n(t) > 0\,, \text{forsome} k > \alpha n\}\,, \tag{32}$$

where $\{C^n(t)\}$ is the rescaled stochastic process defined its generator in Eq. (30). It is known that for the stochastic Smoluchowsky's coagulation model (see Jeon 1998; Eibeck and Wagner 2001; Fournier and Giet 2004; Fournier and Laurençot 2009; Rezakhanlou 2013; Wagner 2005), and for specific coagulation kernel, the sequence

of first passage time (32) has a finite (zero or positive) limit as $n \to \infty$, and that the limit is linked to the gelation (loss of mass) in the deterministic Smoluchowsky's coagulation model. According to the longtime behavior theory for the deterministic Becker–Döring model, it is to be expected that for the SBD, such first passage time (32) can only have infinite limit. However, rate of divergence and summary statistics (mean, variance) as $n \to \infty$ are important open questions.

Acknowledgements E. Hingant thanks the financial support of CAPES/IMPA Brazil during the post-doc at the Universidade Federal de Campina Grande (Paraiba). R. Yvinec thanks the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Stochastic Dynamical Systems in Biology: Numerical Methods and Applications, where part of this work was undertaken.

References

D.J. Aldous, Deterministic and stochastic models for coalescence (aggregation and coagulation): a review of the mean-field theory for probabilists. Bernoulli **5**(1), 3–48 (1999)

M.-T. Alvarez-Martinez, P. Fontes, V. Zomosa-Signoret, J.-D. Arnaud, E. Hingant, L. Pujo-Menjouet, J.-P. Liautard, Dynamics of polymerization shed light on the mechanisms that lead to multiple amyloid structures of the prion protein. Biochim. Biophys. Acta Protein Proteomics **1814**(10), 1305–1317 (2011)

D.F. Anderson, G. Craciun, T.G. Kurtz, Product-form stationary distributions for deficiency zero chemical reaction networks. Bull. Math. Biol. **72**(8), 1947–1970 (2010)

J.M. Ball, J. Carr, Asymptotic behaviour of solutions to the Becker-Döring equations for arbitrary initial data. Proc. R. Soc. Edinb. Sect. A **108**(1–2), 109–116 (1988)

J.M. Ball, J. Carr, O. Penrose, The Becker-Döring cluster equations: basic properties and asymptotic behaviour of solutions. Commun. Math. Phys. **104**(4), 657–692 (1986)

R. Becker, W. Döring, Kinetische Behandlung der Keimbildung in Übersättigten Dämpfen. Ann. Phys. **416**(8), 719–752 (1935)

J.P. Bell, S.N. Burris, Asymptotics for logical limit laws: when the growth of the components is in an RT class. Trans. Am. Math. Soc. **355**(9), 3777–3794 (2003)

E. Ben-Naim, P.L. Krapivsky, Exchange-driven growth. Phys. Rev. E **68**(3), 031104 (2003)

J. Bertoin, *Random Fragmentation and Coagulation Processes*. Cambridge Studies in Advanced Mathematics, vol. 102 (Cambridge University Press, Cambridge, 2006)

A. Bhakta, E. Ruckenstein, Ostwald ripening: a stochastic approach. J. Chem. Phys. **103**(16), 7120 (1995)

J.S. Bhatt, I.J. Ford, Kinetics of heterogeneous nucleation for low mean cluster populations. J. Chem. Phys. **118**(3166), 3166–3166 (2003)

A. Bovier, F. den Hollander, C. Spitoni, Homogeneous nucleation for Glauber and Kawasaki dynamics in large volumes at low temperatures. Ann. Probab. **38**(2), 661–713 (2010)

P.C. Bressloff, Aggregation-fragmentation model of vesicular transport in neurons. J. Phys. A **49**(14), 145601–145616 (2016)

Z. Budrikis, G. Costantini, C.A. La Porta, S. Zapperi, Protein accumulation in the endoplasmic reticulum as a non-equilibrium phase transition. Nat. Commun. **5**, 3620 (2014)

J.J. Burton, Nucleation theory, in *Statistical Mechanics: Part A: Equilibrium Techniques*, ed. by B.J. Berne. Modern Theoretical Chemistry, vol. 5 (Springer, Boston, MA, 1977), pp. 195–234

J.A. Cañizo, Convergence to equilibrium for the discrete coagulation-fragmentation equations with detailed balance. J. Stat. Phys. **129**(1), 1–26 (2007)

J.A. Cañizo, B. Lods, Exponential convergence to equilibrium for subcritical solutions of the Becker-Döring equations. J. Differ. Equ. **255**(5), 905–950 (2013)
J.A. Cañizo, A. Einav, B. Lods, Trend to equilibrium for the Becker-Döring equations: an analogue of Cercignani's conjecture. 1–41 (2015). arXiv:1509.07631
J. Carr, F.P. da Costa, Asymptotic behavior of solutions to the coagulation-fragmentation equations. II. Weak fragmentation. J. Stat. Phys. **77**(1–2), 89–123 (1994)
J. Carr, R.M. Dunwell, Asymptotic behaviour of solutions to the Becker-Döring equations. Proc. Edinb. Math. Soc. (2) **42**(2), 415–424 (1999)
J. Carr, D.B. Duncan, C.H. Walshaw, Numerical approximation of a metastable system. IMA J. Numer. Anal. **15**(4), 505–521 (1995)
C. Cercignani, *Mathematical Methods in Kinetic Theory*, 2nd edn. (Plenum, New York, 1990)
Y.-X. Chau, C. Connaughton, S. Grosskinsky, Explosive condensation in symmetric mass transport models. J. Stat. Mech. Theory E **2015**(11), P11031 (2015)
J.F. Collet, Some modelling issues in the theory of fragmentation-coagulation systems. Commun. Math. Sci. **2**(suppl. 1), 35–54 (2004)
J.F. Collet, T. Goudon, On solutions of the Lifshitz-Slyozov model. Nonlinearity **13**(4), 1239–1262 (2000)
J.F. Collet, T. Goudon, F. Poupaud, A. Vasseur, The Beker-Döring system and its Lifshitz-Slyozov limit. SIAM J. Appl. Math. **62**(5), 1488–1500 (2002)
J.G. Conlon, M. Dabkowski, J. Wu, On large time behavior and selection principle for a diffusive Carr-Penrose model. J. Nonlinear Sci. **26**(2), 453–518 (2016)
P.V. Coveney, J.A.D. Wattis, Analysis of a generalized becker-doring model of self-reproducing micelles. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **452**(1952), 2079–2102 (1996)
J.K. Davis, S.S. Sindi, Initial condition of stochastic self-assembly. Phys. Rev. E **93**(2), 022109 (2016)
P. Dehghanpour, R.H. Schonmann, Metropolis dynamics relaxation via nucleation and growth. Commun. Math. Phys. **188**(1), 89–119 (1997)
F. den Hollander, E. Olivieri, E. Scoppola, Metastability and nucleation for conservative dynamics. J. Math. Phys. **41**(3), 1424–1498 (2000)
J. Deschamps, E. Hingant, E. Yvinec, Quasi steady state approximation of the small clusters in Becker-Döring equations leads to boundary conditions in the Lifshitz-Slyozov limit. Commun. Math. Sci. **15**(5), 1353–1384 (2017)
M.R. D'Orsogna, G. Lakatos, T. Chou, Stochastic self-assembly of incommensurate clusters. J. Chem. Phys. **136**(8), 084110 (2012)
M. Doumic, S. Eugene, P. Robert, Asymptotics of stochastic protein assembly models. 1–19 (2016). arXiv:1603.06335
D.B. Duncan, R.M. Dunwell, Metastability in the classical truncated Becker-Döring equations. Proc. Edinb. Math. Soc. (2) **45**(3), 701–716 (2002)
D.B. Duncan, A.R. Soheili, Approximating the Becker-Döring cluster equations. Appl. Numer. Math. **37**(1–2), 1–29 (2001)
R.M. Dunwell, The Becker-Doring cluster equations. Ph.D. thesis, Heriot-Watt University, Edinburgh (1997)
R. Durrett, B.L. Granovsky, S. Gueron, The equilibrium behavior of reversible coagulation-fragmentation processes. J. Theoret. Probab. **12**(2), 447–474 (1999)
K. Eden, R. Morris, J. Gillam, C.E. MacPhee, R.J. Allen, Competition between primary nucleation and autocatalysis in amyloid fibril self-assembly. Biophys. J. **108**(3), 632–643 (2015)
A. Eibeck, W. Wagner, Stochastic particle approximations for Smoluchowski's coagulation equation. Ann. Appl. Probab. **11**(4), 1137–1165 (2001)
N.M. Ercolani, S. Jansen, D. Ueltschi, Random partitions in statistical mechanics. Electron. J. Probab. **19** (2014)
M.M. Erlihson, B.L. Granovsky, Limit shapes of Gibbs distributions on the set of integer partitions: the expansive case. Ann. Inst. Henri Poincaré Probab. Stat. **44**(5), 915–945 (2008)
M. Escobedo, P. Laurençot, S. Mischler, B. Perthame, Gelation and mass conservation in coagulation-fragmentation models. J. Differ. Equ. **195**(1), 143–174 (2003)

S. Eugene, W.-F. Xue, P. Robert, M. Doumic-Jauffret, Insights into the variability of nucleated amyloid polymerization by a minimalistic model of stochastic protein assembly. J. Chem. Phys. **144**(17), 175101 (2016)
P. Flajolet, R. Sedgewick, *Analytic Combinatorics* (Cambridge University Press, Cambridge, 2009)
N. Fournier, J.-S. Giet, Convergence of the Marcus–Lushnikov process. Methodol. Comput. Appl. Probab. **6**(2), 219–231 (2004)
N. Fournier, P. Laurençot. Marcus–Lushnikov processes, Smoluchowski's and Flory's models. Stoch. Process. Appl. **119**(1), 167–189 (2009)
G.A. Freiman, B.L. Granovsky, Asymptotic formula for a partition function of reversible coagulation-fragmentation processes. Isr. J. Math. **130**(1), 259–279 (2002)
G.A. Freiman, B.L. Granovsky, Clustering in coagulation-fragmentation processes, random combinatorial structures and additive number systems: asymptotic formulae and limiting laws. Trans. Am. Math. Soc. **357**(6), 2483–2507 (2005)
B.L. Granovsky, Asymptotics of counts of small components in random structures and models of coagulation-fragmentation. ESAIM Probab. Stat. **17**, 531–549 (2013)
D. Han, X.S. Zhang, W.A. Zheng, Subcritical, critical and supercritical size distributions in random coagulation-fragmentation processes. Acta Math. Sin. (Engl. Ser.) **24**(1), 121–138 (2008)
S. Hariz, J.F. Collet, A modified version of the Lifshitz-Slyozov model. Appl. Math. Lett. **12**(1), 81–85 (1999)
N. Hoze, D. Holcman, Coagulation-fragmentation for a finite number of particles and application to telomere clustering in the yeast nucleus. Phys. Lett. A **376**(6), 845–849 (2012)
N. Hoze, D. Holcman, Modeling capsid kinetics assembly from the steady state distribution of multi-sizes aggregates. Phys. Lett. A **378**(5), 531–534 (2014)
N. Hoze, D. Holcman, Kinetics of aggregation with a finite number of particles and application to viral capsid assembly. J. Math. Biol. **70**, 1685–1705 (2015)
J. Hu, H.G. Othmer, A theoretical analysis of filament length fluctuations in actin and other polymers. J. Math. Biol. **63**, 1001–1049 (2011)
P.E. Jabin, B. Niethammer, On the rate of convergence to equilibrium in the Becker-Döring equations. J. Differ. Equ. **191**(2), 518–543 (2003)
I. Jeon, Existence of gelling solutions for coagulation-fragmentation equations. Commun. Math. Phys. **567**, 541–567 (1998)
J. Johansson, Stochastic analysis of nucleation rates. Phys. Rev. E **93**(2), 022801 (2016)
F.P. Kelly, *Reversibility and Stochastic Networks* (Cambridge University Press, Cambridge, 1979)
A.Y. Khinchin, *Mathematical Foundations of Quantum Statistics*, ed. by I. Shapiro. Translation from the first (1951) Russian edition (Graylock Press, Albany, NY, 1960)
J.R. King, J.A.D. Wattis, Asymptotic solutions of the Becker-Döring equations with size-dependent rate constants. J. Phys. A **35**(6), 1357–1380 (2002)
M. Kreer, Classical Becker-Döring cluster equations: rigorous results on metastability and long-time behaviour. Ann. Phys. (8) **2**(4), 398–417 (1993)
P. Laurençot, Weak solutions to the Lifshitz-Slyozov-Wagner equation. Indiana Univ. Math. J. **50**(3), 1319–1346 (2001)
P. Laurençot, The discrete coagulation equations with multiple fragmentation. Proc. Edinb. Math. Soc. (2) **45**(1), 67–82 (2002)
P. Laurençot, S. Mischler, From the Becker-Döring to the Lifshitz-Slyozov-Wagner equations. J. Stat. Phys. **106**(5–6), 957–991 (2002)
P. Laurençot, D. Wrzosek, The Becker-Döring model with diffusion. I. Basic properties of solutions. Colloq. Math. **75**(2), 245–269 (1998)
I.M. Lifshitz, V.V. Slyozov, The kinetics of precipitation from supersaturated solid solutions. J. Phys. Chem. Solids **19**(1–2), 5–50 (1961)
B. Linse, S. Linse, Monte carlo simulations of protein amyloid formation reveal origin of sigmoidal aggregation kinetics. Mol. Biosyst. **7**, 2296–2303 (2011)
A.A. Lushnikov, Coagulation in finite systems. J. Colloid Interf. Sci. **65**(2), 276–285 (1978)
A.H. Marcus, Stochastic coalescence. Technometrics **10**(1), 133–143 (1968)

B. Niethammer, On the evolution of large clusters in the Becker-Döring model. J. Nonlinear Sci. **13**(1), 115–155 (2003)
B. Niethammer, A scaling limit of the Becker-Döring equations in the regime of small excess density. J. Nonlinear Sci. **14**(5), 453–468 (2004)
B. Niethammer, Effective theories for Ostwald ripening, in *Analysis and Stochastics of Growth Processes and Interface Models*, ed. by P. Mörters, R. Moser, M. Penrose, H. Schwetlick, J. Zimmer (Oxford University Press, Oxford, 2008), pp. 223–243
B. Niethammer, R.L. Pego, On the initial-value problem in the Lifshitz-Slyozov-Wagner theory of Ostwald ripening. SIAM J. Math. Anal. **31**(3), 467–485 (2000)
B. Niethammer, F. Otto, J.J.L. Velázquez, On the effect of correlations, fluctuations and collisions in Ostwald ripening. in *Analysis, Modeling and Simulation of Multiscale Problems*, ed. by A. Mielke (Springer, Berlin, 2006), pp. 501–530
O. Penrose, Metastable states for the Becker-Döring cluster equations. Commun. Math. Phys. **124**(4), 515–541 (1989)
O. Penrose, The Becker-Döring equations for the kinetics of phase transitions. Lecture Notes at Strathclyde University, pp. 1–12 (1995)
O. Penrose, The Becker-Döring equations at large times and their connection with the LSW theory of coarsening. J. Stat. Phys. **89**(1–2), 305–320 (1997). Dedicated to Bernard Jancovici
O. Penrose, Nucleation and droplet growth as a stochastic process, in *Analysis and Stochastics of Growth Processes and Interface Models* (Oxford University Press, Oxford, 2008), pp. 1–12
O. Penrose, A. Buhagiar, Kinetics of nucleation in a lattice gas model: microscopic theory and simulation compared. J. Stat. Phys. **30**(1), 219–241 (1983)
O. Penrose, J.L. Lebowitz, Towards a rigorous molecular theory of metastability, in *Fluctuation Phenomena*, ed. by E.W. Montroll, J.L. Lebowitz. Studies in Statistical Mechanics, vol. 7 (Elsevier, Amsterdam, 1979), pp. 293–340
O. Penrose, J.L. Lebowitz, J. Marro, M.H. Kalos, A. Sur, Growth of clusters in a first-order phase transition. J. Stat. Phys. **19**(3), 243–267 (1978)
S. Prigent, A. Ballesta, F. Charles, N. Lenuzza, P. Gabriel, L.M. Tine, H. Rezaei, M. Doumic, An efficient kinetic model for assemblies of amyloid fibrils and its application to polyglutamine aggregation. PloS One **7**(11), e43273–e43273 (2012)
F. Rezakhanlou, Gelation for Marcus–Lushnikov process. Ann. Probab. **41**(3), 1806–1830 (2013)
H.G. Rotstein, Cluster-size dynamics: a phenomenological model for the interaction between coagulation and fragmentation processes. J. Chem. Phys. **142**(22), 224101 (2015)
J.W.P. Schmelzer (ed.), *Nucleation Theory and Applications* (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2005)
F. Schweitzler, L. Schimansky-Geier, W. Ebeling, H. Ulbricht, A stochastic approach to nucleation in finite systems: theory and computer simulations. Phys. A **150**, 261–279 (1988)
M. Slemrod, Trend to equilibrium in the Becker-Döring cluster equations. Nonlinearity **2**(3), 429–443 (1989)
M. Slemrod, The Becker-Döring equations, in *Modeling in Applied Sciences: A Kinetic Theory Approach*, ed. by N. Bellomo, M. Pulvirenti. Modeling and Simulation in Science, Engineering and Technology (Birkhäuser, Boston, MA, 2000), pp. 149–171
V. Ssemaganda, G. Warnecke, Existence of metastable solutions for a thermodynamically consistent Becker-Döring model. J. Appl. Anal. **19**(1), 91–124 (2013)
J.J.L. Velázquez, The Becker-Döring equations and the Lifshitz-Slyozov theory of coarsening. J. Stat. Phys. **92**(1–2), 195–236 (1998)
J.J.L. Velázquez, On the effect of stochastic fluctuations in the dynamics of the Lifshitz-Slyozov-Wagner model. J. Stat. Phys. **99**(1–2), 57–113 (2000)
W. Wagner, Explosion phenomena in stochastic coagulation–fragmentation models. Ann. Appl. Probab. **15**(3), 2081–2112 (2005)
J.A.D. Wattis, An introduction to mathematical models of coagulation-fragmentation processes: a discrete deterministic mean-field approach. Phys. D **222**(1–2), 1–20 (2006)
J.A.D. Wattis, C.D. Bolton, P.V. Coveney, The Becker-Döring equations with exponentially size-dependent rate coefficients. J. Phys. A **37**(8), 2895–2912 (2004)

P. Whittle, Statistical processes of aggregation and polymerization. Math. Proc. Camb. Philos. Soc. **61**(2), 475–495 (1965)

R. Yvinec, M.R. D'Orsogna, T. Chou, First passage times in homogeneous nucleation and self-assembly. J. Chem. Phys. **137**(24), 244107 (2012)

R. Yvinec, S. Bernard, E. Hingant, L. Pujo-Menjouet, First passage times in homogeneous nucleation: dependence on the total number of particles. J. Chem. Phys. **144**, 34106 (2016)

Coagulation-Fragmentation with a Finite Number of Particles: Models, Stochastic Analysis, and Applications to Telomere Clustering and Viral Capsid Assembly

Nathanael Hoze and David Holcman

1 Introduction

Clustering processes are generic in statistical physics and biology. For example, in astrophysics, masses can form aggregate under the gravitation force, while in biochemistry, molecules interact to form colloids that aggregate in solution [7]. In cell biology, aggregation underlies beta-amyloid structure formation involved in Alzheimer's disease or chromosomal organization in the cell nucleus. However, a new class of mathematical problems appears with the need to analyze clustering with a finite number of random particles such as the organization of the chromosome ends [14] or viral capsid assembly in cells. These processes are modeled as coagulation-fragmentation.

Irreversible aggregation of many particles in clusters was already described by Von Smoluchowski in 1916 [34] to model an infinite number of interacting molecules. When a cluster can lose or gain only one particle at a time, the Smoluchowski equations become the Becker-Döring model which consists in an ensemble of coagulation-fragmentation equations [4, 7, 23, 35]. Nowadays, determinist, stochastic, asymptotic, and numerical methods are developed to study steady-state and transient properties of clustering based on molecular components [2, 8, 9, 26, 33]. Another class of problem concerns the clustering with an infinite number of particles (Marcus-Lushnikov process) [21, 24, 25], but much less is

N. Hoze
Institut für Integrative Biologie, ETH, Universitätstrasse 16, 8092 Zürich, Switzerland

D. Holcman (✉)
Institute for Biology École Normale Supérieure, Applied Mathematics and Computational Biology, Paris, France

Churchill College, University of Cambridge, Storey's Way, Cambridge CB3 0DS, UK
e-mail: david.holcman@ens.fr

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_10

known about coagulation-fragmentation with a finite number of particles [11]. When the cluster size cannot exceed a given threshold, new difficulties arise in the analysis of the coagulation-fragmentation equations [14, 36]. These models are relevant in molecular genetics for characterizing the organization of the chromosome ends [14] or to model viral capsid assembly in cell biology [16, 17, 37].

We review here several models, asymptotic and combinatorial results as well as a generalization of the Gillespie's algorithm [15] to study aggregation in spatially inhomogeneous environment. In the first section, we describe the Smoluchowski equations for coagulation-fragmentation. In the second, we present a general analysis and result about clustering with a finite number of particle. Section 3 is dedicated to Gillespie's algorithm in spatially inhomogeneous environment, applied to telomere organization in yeast. In Sects. 4 and 5, we present asymptotic methods for capsid viral assembly and the analysis of single particle trajectories.

2 Primer in Smoluchowski Equations for Coagulation-Fragmentation

2.1 Coagulation-Fragmentation with an Infinite Number of Particles

This section summarizes the Smoluchowski equations for coagulation-fragmentation that consist of an infinite system of differential equations for the number $n(j,t)$ of clusters of size j at time t in a population of infinite size [34]. The coagulation process is characterized by the rate $C(i,j)$ by which two clusters of size i and j coalesce to form a cluster of size $i+j$, while fragmentation with rate $F(i,j)$ describes that a cluster of size $i+j$ dissociates into a cluster of size i and a cluster of size j. The conservation of mass equation is given by

$$\begin{aligned}\frac{dn(j,t)}{dt} &= \frac{1}{2}\sum_{k=0}^{j-1} C(k,j-k)n(k,t)n(j-k,t) - n(j,t)\sum_{k=1}^{\infty} C(j,k)n(k,t) \\ &\quad -n(j,t)\sum_{k=1}^{j-1} F(k,j-k) + \sum_{k=1}^{\infty} F(j,k)n(k+j,t), \end{aligned} \tag{1}$$

where the index j can take values between 1 and ∞ and the first line in the left-hand side corresponds to the coagulation and the second accounts for the fragmentation. This system of equations is a mean-field deterministic model of the coagulation-fragmentation process that do not describe intrinsic cluster interactions.

Coagulation-fragmentation processes (CFP) satisfies the balance condition [10], for which there exists a function $a(i) = a_i$ such that $\forall i,j \in \mathbb{N}$

$$\frac{C(i,j)}{F(i,j)} = \frac{a(i+j)}{a(i)a(j)}. \tag{2}$$

When the total number of clusters is fixed, the probability distribution function of the number of clusters can be computed, as well as the probability distribution that the number of cluster of size i is m_i so that the distribution of sizes of clusters is $(m_1, \ldots, m_n)$. When there are exactly N particles and the total number of clusters is fixed to K, the following identity for number conservation is satisfied [22]:

$$\sum_{i=1}^{N} m_i = K. \tag{3}$$

When the total number of clusters is K, the conditional probability distribution function is given by

$$p'(m_1, \ldots, m_N | K) = \frac{1}{C_{N,K}} \frac{a(1)^{m_1} \ldots a(N)^{m_N}}{m_1! \ldots m_N!},$$

where the normalization constant $C_{N,K}$ will be described below (see formula (53)). These formulas are used to compute the statistical moments for the cluster distributions.

2.2 Continuous-Time Markov Chain Equations for a Finite Number of Particles

The steady-state distribution for a CFP stochastic model with a finite number of N particles is described by a continuous-time Markov chain equations in the cluster configuration space. We start with N particles distributed in clusters of size $(n_1, \ldots, n_K)$ that can undergo coagulation or fragmentation events under the constraint that

$$\sum_{k=1}^{K} n_k = N. \tag{4}$$

To study the distribution of particles in clusters, we use the decomposition of the integer N in a sum of positive integers (integer partition) [3]. The partitions of the integer N are described in dimension N by the ensemble

$$P_N = \left\{ (n_1, \ldots, n_N) \in \mathbb{N}^N; \sum_{i=1}^{N} n_i = N \text{ and } n_1 \geq \cdots \geq n_N \geq 0 \right\}. \tag{5}$$

The probability $P(n_1, \ldots, n_N, t)$ of the configuration $(n_1, \ldots, n_N)$ at time t satisfies an ensemble of close equations obtained by considering all possible coagulations or fragmentations between time t and $t + \Delta t$:

- Two clusters of size n_i and n_j coagulate with a probability $C(n_i, n_j)\Delta t$ to form a cluster of size $n_i + n_j$.
- A cluster of size n_i dissociates into two clusters of size k and $n_i - k$ with a probability $F(k, n_i - k)\Delta t$.
- Nothing happens with the probability $1 - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} C(n_i, n_j)\Delta t - \sum_{i=1}^{N}\sum_{k=1}^{n_i-1} F(k, n_i - k)\Delta t$.

Thus, the probability $P(n_1, \ldots, n_N, t)$ satisfies

$$\begin{aligned}\frac{d}{dt}P(n_1,\ldots,n_N,t) = &-\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} C(n_i,n_j) + \sum_{i=1}^{N}\sum_{k=1}^{n_i-1} F(k,n_i-k)\right)P(n_1,\ldots,n_N,t)\\ &+\sum_{k=1}^{N}\sum_{\substack{n_i'>0,n_j'>0\\ n_i'+n_j'=n_k}} C(n_i',n_j')P(n_1,\ldots,n_i',\ldots,n_j',\ldots,n_N,t)\\ &+\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} F(n_i,n_j)P(n_1,\ldots,n_i+n_j,\ldots,n_N,t). \qquad (6)\end{aligned}$$

Moreover, $C(n_i, n_j) = 0$ if either n_i or n_j is equal to 0. The partitions of the integer N are described by the set

$$P_N = \left\{(n_1,\ldots,n_N)\in\mathbb{N}^N; \sum_{i=1}^{N} n_i = N \text{ and } n_1 \geq \cdots \geq n_N \geq 0\right\} \qquad (7)$$

and the ensemble of decompositions

$$P'_N = \left\{(m_1,\ldots,m_N)\in\mathbb{N}^N; \sum_{i=1}^{N} i m_i = N \text{ and } n_1,\ldots,n_N \geq 0\right\}. \qquad (8)$$

In the ensemble P'_N, m_i is the number of occurrence of integer i in the partition of the integer N. The two ensembles P_N and P'_N correspond to different representations of the clusters distributions.

For example, $N = 9$ particles are distributed in two clusters of one particle, two clusters of two, and one cluster of three and the distribution is $(3, 2, 2, 1, 1, 0, 0, 0, 0) \in P_9$, and $(2, 2, 1, 0, 0, 0, 0, 0, 0) \in P'_9$.

When the coefficient C and F satisfies relation (2), there exists an invariant measure [10] for the steady-state probability of a given configuration $(m_1, \ldots, m_N) \in P'_N$, given by

$$P'(m_1,\ldots,m_N) = \frac{1}{C_N}\frac{a_1^{m_1}\ldots a_N^{m_N}}{m_1!\ldots m_N!}, \qquad (9)$$

where C_N is a normalization constant. Computing the normalization constant explicitly is difficult [32]. In the next subsection, we estimate the probability of occurrence of a certain cluster configuration $(m_1, \ldots, m_N)$.

2.3 Description of the Cluster Partitions with a Finite Number of Particles

To determine the cluster distribution at equilibrium, we compute here the probability of a configuration when the number of clusters K is fixed. We also find the probability of having K clusters. The number of distributions of N particles into K clusters is the cardinal of the ensemble

$$P_{N,K} = \left\{ (n_1, \ldots, n_K) \in (\mathbb{N})^N; \sum_{i=1}^{K} n_i = N \text{ and } n_1 \geq \cdots \geq n_N \geq 0 \right\}, \tag{10}$$

which is also the ensemble of the partitions of the integer N as a sum of K integers. This ensemble is in bijection with

$$P'_{N,K} = \left\{ (m_1, \ldots, m_N) \in \mathbb{N}^N; \sum_{i=1}^{N} i m_i = N \text{ and } \sum_{i=1}^{N} m_i = K \right\}, \tag{11}$$

where the application $P_{N,K} \to P'_{N,K}$ defined by

$$(n_1, \ldots, n_K) \mapsto (m_1, \ldots, m_N) = \left(\sum_{i=1}^{K} 1_{\{n_i=1\}}, \ldots, \sum_{i=1}^{K} 1_{\{n_i=N\}} \right) \tag{12}$$

maps the partition $(n_1, \ldots, n_N)$ where N is written as a sum of K positive integers to the number of occurrence of each integer into the image partition. The partitions of N are written as

$$P_N = \bigcup_K P_{N,K} \text{ and } P'_N = \bigcup_K P'_{N,K}. \tag{13}$$

In Sects. 3.1–3.3, we derive explicitly expressions for the probabilities of configurations in $P'_{N,K}$.

2.4 Statistical Moments for the Cluster Configurations When the Number of Clusters Is Fixed

The probability of a configuration $(m_1, \ldots, m_N)$, when the total number of clusters is equal to K, is

$$p'(m_1, \ldots, m_N | K) = \frac{\frac{a_1^{m_1} \ldots a_N^{m_N}}{m_1! \ldots m_N!}}{C_{N,K}}. \tag{14}$$

where

$$C_{N,K} = \sum_{(m_i) \in P'_{N,K}} \frac{a_1^{m_1} \ldots a_N^{m_N}}{m_1! \ldots m_N!}. \tag{15}$$

The normalization factor of Eq. (14) is computed using the partial sums

$$S_N(x) = \sum_{i=1}^{N} a_i x^i. \tag{16}$$

The functions S^K and S_N^K have the same Nth order coefficient and this coefficient determines $C_{N,K}$. We recall [18] the

Theorem 2.1 *When the number of clusters is equal to K for a total of N particles, the mean number of clusters of size i is*

$$\langle M_i \rangle_{N,K} = a_i \frac{C_{N-i,K-1}}{C_{N,K}}, \tag{17}$$

where a_i and $C_{N,K}$ are defined in (2) and (15), respectively.

Furthermore, $\langle M_i \rangle_{N,K} = 0$ if $i > N - K + 1$. Interestingly,

Theorem 2.2 *The second moment of the number of clusters of size i is*

$$\begin{aligned} \langle M_i^2 \rangle_{N,K} &= \frac{1}{C_{N,K}} \sum_{P'_{N,K}} m_i^2 \frac{a_1^{m_1} \ldots a_N^{m_N}}{m_1! \ldots m_N!} \\ &= a_i^2 \frac{C_{N-2i,K-2}}{C_{N,K}} + a_i \frac{C_{N-i,K-1}}{C_{N,K}}, \end{aligned} \tag{18}$$

and the covariance is

$$\langle M_{i,j}^2 \rangle_{N,K} - \langle M_j^2 \rangle_{N,K} \langle M_j \rangle_{N,K} = a_i a_j \left(\frac{C_{N-i-j,K-2}}{C_{N,K}} - \frac{C_{N-i,K-1} C_{N-j,K-1}}{C_{N,K}^2} \right). \tag{19}$$

The proofs can be found in [18].

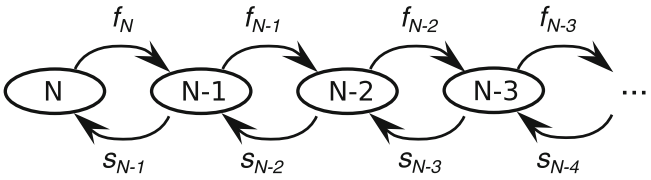


Fig. 1 Markov chain representation for the number of clusters. s_K (respectively, f_K) is the separation (respectively, formation) rate of a cluster when there are K clusters

2.5 Distribution of the Number of Clusters

In the previous section, we introduce the probability distribution of a cluster configuration and the statistical moments for a fixed number of clusters. In this section, we describe the statistics of the entire cluster configurations using the probability distribution of the *number of clusters*. Our goal is to study the time dependent probability density function

$$P_K(t) = P\{K \text{ clusters at time } t\}, \tag{20}$$

which is associated with a birth-and-death process: the probability of having K clusters at time $t + \Delta t$ is the sum of the probability of starting at time t with $K - 1$ clusters and one of them dissociates into two smaller ones plus the probability of starting with $K+1$ clusters and two of them associate plus the probability of starting with K and nothing happens (Fig. 1).

The first probability is the product of P_{K-1} by the transition rate $s_{K-1}\Delta t$ to go from state with $K - 1$ clusters to K, while the second is the transition from $K + 1$ to K, which is the product of P_{K+1} by the transition rate $f_{K+1}\Delta t$ of going from $K + 1$ clusters to K. The master equations are given by

$$\begin{cases} \dot{P}_1(t) = -s_1 P_1(t) + f_2 P_2(t) \\ \dot{P}_K(t) = -(f_K + s_K)P_K(t) + f_{K+1}P_{K+1}(t) + s_{K-1}P_{K-1}(t) \\ \dot{P}_N(t) = -f_N P_N(t) + s_{N-1}P_{N-1}(t). \end{cases} \tag{21}$$

The steady probability is defined by

$$\Pi_K = \lim_{t\to\infty} P_K(t) \tag{22}$$

where there are K clusters at steady state. The steady-state probabilities of the number of clusters are solutions of the system

$$\begin{cases} 0 = -s_1 \Pi_1 + f_2 \Pi_2 \\ 0 = -(f_K + s_K)\Pi_K + f_{K+1}\Pi_{K+1} + s_{K-1}\Pi_{K-1} \\ 0 = -f_N \Pi_N + s_{N-1}\Pi_{N-1}, \end{cases} \tag{23}$$

with the normalization condition

$$\sum_{K=1}^{N} \Pi_K = 1. \tag{24}$$

The probabilities Π_K are given by the ratio

$$\frac{\Pi_K}{\Pi_{K-1}} = \frac{s_{K-1}}{f_K} \text{ for } K \geq 2 \tag{25}$$

and the coefficients s_K and f_K are the mean-field separation and formation rates, respectively. Whereas the cluster configurations when the number of clusters is fixed depend only on the kernel a_i, the statistics of the number of clusters depends on the cluster fragmentation and coagulation rates F and C.

In the following, we will focus on the coagulation condition $C(i,j) = 1$ and the fragmentation $F(i,j) = \frac{a_i a_j}{a_{i+j}}$ to state the

Theorem 2.3 *When $C(i,j) = 1$ and $F(i,j) = \frac{a_i a_j}{a_{i+j}}$, the separation rate when there are K clusters is given by*

$$s_K = \frac{\sum_{i=1}^{N} \sum_{j=1}^{i-1} a_j a_{i-j} C_{N-i,K-1}}{C_{N,K}} \tag{26}$$

and the formation rate when there are K clusters is

$$f_K = \frac{K(K-1)}{2}. \tag{27}$$

Using these results, we can now describe the statistics of the entire cluster configurations. Using Bayes rule, the probability of a configuration $(m_1, \ldots, m_N)$ that contains K clusters is the product of the conditional probability $p'(m_1, \ldots, m_N | K)$ by the probability of having K clusters

$$p'(m_1, \ldots, m_N, K) = p'(m_1, \ldots, m_N | K) \Pi_K. \tag{28}$$

The mean number of clusters of size i is thus

$$\langle M_i \rangle_N = \sum_{K=1}^{N} \Pi_K \langle M_i \rangle_{N,K}. \tag{29}$$

2.6 The Probability to Find Two Particles in the Same Cluster

When the mean number of clusters has reached its equilibrium, particles can still be exchanged between clusters. This exchange is characterized by the probability to find two particles in the same cluster.

When the distribution of the clusters is $(n_1, \ldots, n_K)$, the probability $P_2(n_1, \ldots, n_K)$ to find two given particles in the same cluster is obtained by using the probability to choose the first particle in the cluster n_i, which is equal to the number of particles in the cluster divided by the total number of particles $\frac{n_i}{N}$. The probability to have the second particle in the same cluster is $\frac{n_i-1}{N-1}$. Summing over all possibilities, we get

$$P_2(n_1, \ldots, n_K) = \sum_{i=1}^{K} \frac{n_i}{N} \frac{n_i - 1}{N-1} = \frac{1}{N(N-1)} \left(\sum_{i=1}^{K} n_i^2 - N \right). \tag{30}$$

We note that

$$\sum_{j=1}^{K} n_j^2 = \sum_{i=1}^{N} i^2 m_i, \tag{31}$$

thus we get

$$\begin{aligned} \sum_{(n_1,\ldots,n_K)\in P_{N,K}} p(n_1, \ldots, n_K) \sum_{j=1}^{K} n_j^2 &= \sum_{(m_i)\in P'_{N,K}} p(m_i) \sum_{j=1}^{N} j^2 m_j \\ &= \sum_{j=1}^{N} j^2 \sum_{(m_i)\in P'_{N,K}} m_j p(m_i) \\ &= \sum_{j=1}^{N} j^2 \langle M_j \rangle_{N,K}, \end{aligned} \tag{32}$$

where $\langle M_j \rangle_{N,K}$ is the mean number of clusters of size j, when there are N particles distributed in K clusters Eq. (17). Taking into account all possible distributions of clusters, we obtain that the probability $\langle P_2 \rangle$ to find two particles in the same cluster is

$$\langle P_2 \rangle = \sum_{K=1}^{N} \sum_{(n_1,\ldots,n_K)\in P_{N,K}} P_2(n_1, \ldots, n_K) p(n_i) \Pi_K, \tag{33}$$

which can be written, using expressions (30) and (33) as

$$\langle P_2 \rangle = \frac{1}{N(N-1)} \sum_{K=1}^{N} \Pi_K \sum_{j=1}^{N} j^2 \langle M_j \rangle_{N,K} - \frac{1}{N-1}. \tag{34}$$

This approach can be generalized to the probability of having $n \geq 2$ particles together.

3 Examples of Coagulation-Fragmentation with a Finite Number of Particles

We shall now summarize several results in the three examples:

1. $a_i = a$
2. $a_i = a$ for $i < M$ and $a_i = 0$ if $i \geq M$
3. We finally consider the case $a_i = ai$.

3.1 Example 1: The Case $a_i = a$

When $a_i = a$, the separation and formation rates s_K and f_K are computed with $F(i,j) = a$ and $C(i,j) = 1$. A cluster of size n dissociates at a rate $\sum_{i=1}^{n-1} F(i, n-i) = (n-1)a$ and the sizes of the resulting clusters are uniformly distributed between 1 and $n-1$. The total transition rate from a configuration of K to $K+1$ clusters is the sum over all possible dissociation rates

$$s_K = \sum_{i=1}^{K} (n_i - 1)a = (N-K)a. \tag{35}$$

The formation rate is proportional to the number of pairs

$$f_K = \frac{K(K-1)}{2}. \tag{36}$$

The steady-state probability Π_K for the number of clusters of size K satisfies the time independent master equation

$$\begin{cases} s_1 \Pi_1 = f_2 \Pi_2, \\ \mu_1 (f_K + s_K)\Pi_K = f_{K+1}\Pi_{K+1} + s_{K-1}\Pi_{K-1}, \\ f_N \Pi_N = s_{N-1}\Pi_{N-1}, \end{cases} \tag{37}$$

which leads to the relation

$$\Pi_{K+1} = (2a)^K \frac{(N-1)!}{K!(K+1)!(N-K-1)!}\Pi_1. \tag{38}$$

With the normalization condition $\sum_K \Pi_K = 1$, the probability Π_1 is expressed with a hypergeometric series

$$\Pi_1 = \frac{1}{{}_1F_1(-N+1; 2; -2a)}, \tag{39}$$

where

$$ {}_1F_1(a;b;z) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n} \frac{z^n}{n!}, \tag{40} $$

is Kummer's confluent hypergeometric function ([1, pp. 503–535]) and

$$ (x)_n = x(x+1)\ldots(x+n-1) \tag{41} $$

is the Pochhammer symbol. The average number of clusters at steady state is

$$ \begin{aligned} \mu_1(a) &= \sum_{K=1}^{N} K \Pi_K \\ &= \Pi_1 \frac{d}{dz} \left(z {}_1F_1(-N+1;2;z) \right)_{|z=-2a}. \end{aligned} \tag{42} $$

The derivative of the Kummer's function is

$$ \frac{d}{dz} {}_1F_1(a;b;z) = \frac{a}{b} {}_1F_1(a+1;b+1;z). \tag{43} $$

The mean number of clusters is expressed as

$$ \begin{aligned} \mu_1(a) &= 1 + a(N-1) \frac{{}_1F_1(-N+2;3;-2a)}{{}_1F_1(-N+1;2;-2a)}, \\ &= 1 + a(N-1) G_1, \end{aligned} \tag{44} $$

where we note G_1 the function defined by

$$ G_1 = \frac{{}_1F_1(-N+2;3;-2a)}{{}_1F_1(-N+1;2;-2a)}. \tag{45} $$

More generally, we introduce the functions G_i defined by

$$ G_i = \frac{{}_1F_1(-N+1+i;2+i;-2a)}{{}_1F_1(-N+1;2;-2a)}. \tag{46} $$

All moments of the probability distribution Π_K can be computed and the nth-order moment μ_n is expressed using the operator H defined by

$$ H(f)(z) = \frac{d}{dz} z f(z), \tag{47} $$

by

$$\mu_n = \sum_{n=1}^{N} K^n \Pi_K = \frac{H^{(n)}({}_1F_1(-N+1;2;z))_{|z=-2a}}{{}_1F_1(-N+1;2;-2a)}. \tag{48}$$

Using the differentiation formula for the hypergeometric function (43), the moments are

$$\mu_n = \sum_{k=0}^{n} \alpha_k^n \frac{\Pi_{k+1}}{\Pi_1} G_k, \tag{49}$$

where

$$\alpha_k^n = \begin{cases} k! \sum_{j=0}^{k/2} (-1)^j \dfrac{(k+1-j)^n + (j+1)^n}{(k-j)!} & \text{if } k \text{ is even,} \\ k! \sum_{j=0}^{(k-1)/2} (-1)^j \dfrac{(k+1-j)^n - (j+1)^n}{(k-j)!} & \text{if } k \text{ is odd,} \end{cases}$$

and $\alpha_0^n = \alpha_n^n = 1$. The variance of the number of clusters is given by

$$\begin{aligned} \langle V_\infty(a) \rangle &= \mu_2 - \mu_1^2 = a(N-1)G_1(a,N) + \frac{2}{3}a^2(N-1)(N-2)G_2(a,N) \\ &\quad -a^2(N-1)^2 G_1^2(a,N). \end{aligned} \tag{50}$$

3.1.1 Number of Clusters of a Given Size

The statistical moments for the size of clusters are computed from relation (17) and the mean number of clusters of size n when there are K clusters is

$$\langle M_n \rangle_{N,K} = \sum_{(m_i) \in P'_{N,K}} m_n p'(m_i|K) = a \frac{C_{N-n,K-1}}{C_{N,K}}. \tag{51}$$

The normalizing constant $C_{N,K}$ given in Eq. (15) is the Nth order coefficient of S^K, where S is the generating function

$$S(x) = \sum_{i=1}^{\infty} a_i x^i = a \frac{x}{1-x}. \tag{52}$$

The coefficient $C_{N,K}$ is thus equal to the $N-K$th order coefficient of $\frac{1}{K!}\frac{a^K}{(1-x)^K}$. By differentiating $N-K$ times $\frac{1}{(1-x)^K}$ and estimating the derivative at $x=0$. We obtain that

$$C_{N,K} = \frac{a^K}{K!}\frac{(N-1)!}{(K-1)!(N-K)!}. \tag{53}$$

Thus, by combining (51) and (53),

$$\langle M_n\rangle_{N,K} = \frac{(N-n-1)!K!(N-K)!}{(N-1)!(K-2)!(N-n-K+1)!}, \tag{54}$$

The mean number of clusters of size n is obtained by summing over all possibilities configuration with K clusters,

$$\langle M_n\rangle = \sum_{K=1}^{N}\langle M_n\rangle_{N,K}\Pi_K = \frac{(N-n-1)!}{(N-1)!}\sum_K \frac{K(K-1)(N-K)!}{(N-n-K+1)!}\Pi_K.$$

Using expression (38) for Π_K, we obtain

$$\langle M_n\rangle = 2a\frac{{}_1F_1(-N+1+n;2;-2a)}{{}_1F_1(-N+1;2;-2a)} \text{ if } n<N, \tag{55}$$

and

$$\langle M_N\rangle = \frac{1}{{}_1F_1(-N+1;2;-2a)}. \tag{56}$$

The mean number of clusters of size N is exactly equal to the probability $\Pi_1(N)$ of having one cluster when there is N particles [see Eq. (39)].

3.1.2 Probability to Find Two Particles in the Same Cluster

The probability to find two particles in the same cluster for a constant kernel $a_i = a$, when there are N particles, is

$$\langle P_2\rangle = G_1, \tag{57}$$

where G_1 is defined in (45). Indeed the probability that two particles are in the same cluster is

$$\begin{aligned}\langle P_2\rangle &= \frac{1}{N(N-1)}\sum_{K=1}^{N}\Pi_K\sum_{j=1}^{N}j^2\langle M_j\rangle_{N,K} - \frac{1}{N-1}\\ &= \frac{1}{N(N-1)}\sum_{K=1}^{N}\Pi_K\left(N+2N\frac{N-K}{K+1}\right) - \frac{1}{N-1},\end{aligned} \tag{58}$$

where the average number of clusters of size j when there is a total of K clusters is given by relation (54). Thus,

$$\langle P_2 \rangle = \frac{2}{N-1} \sum_{K=1}^{N} \Pi_K \frac{N-K}{K+1} = -\frac{2}{N-1} + 2\frac{N+1}{N-1} \sum_{K=1}^{N} \frac{1}{K+1} \Pi_K. \quad (59)$$

which is the definition of G_1 Eq. (45). For large N, we thus obtain that the probability that two particles are in the same cluster is

$$\langle P_2 \rangle \approx \sqrt{\frac{2}{aN}}. \quad (60)$$

The results presented in this section were used to study the distribution of clusters in biological systems such as telomere organization in yeast [14].

3.2 Example 2: The Case $a_i = a$ for $i < M$ and $a_i = 0$ if $i \geq M$

When N particles can associate or dissociate with a constant rate, but cannot form clusters of more than M particles, the configuration space for the distribution of N particles in K clusters of size less than M is now

$$P'_{N,K,M} = \left\{ (m_i)_{1 \leq i \leq M}; \sum_{i=1}^{M} i m_i = N, \sum_{i=1}^{M} m_i = K \right\}. \quad (61)$$

First, the minimal number of clusters is necessarily bounded by $K \geq N/M$, since the opposite would imply a cluster of at least $M+1$ particles. The probability of a configuration $(m_1, \ldots, m_M) \in P'_{N,K,M}$ is equal to

$$P'\{(m_1, \ldots, m_M) \in P'_{N,K,M}\} = \frac{1}{C_{N,K,M}} \frac{1}{m_1! \ldots m_M!}, \quad (62)$$

where the normalization constant $C_{N,K,M}$ is the Nth order coefficient of

$$(aX + aX^2 + \cdots + aX^M)^K = a^K \frac{1}{(1-X)^K} \sum_{n=0}^{K} \binom{K}{n} (-1)^n X^{nM+K}. \quad (63)$$

Then the Nth order coefficient of the polynomial is obtained by finding the $(N - nM - K)$th order coefficient of $(1-X)^{-K}$

$$C_{N,K,M} = a^K \sum_{n=0}^{K} \binom{K}{n} (-1)^n \frac{1}{(N-(nM+K))!} D^{(N-(nM+K))} \left(\frac{1}{(1-X)^K} \right)_{|X=0}, \quad (64)$$

where we write $D^{(n)}$ the nth order derivative. Thus, setting $K_0 = \lfloor \frac{N-K}{M} \rfloor$, where $\lfloor . \rfloor$ is the floor function, we have

$$C_{N,K,M} = a^K K \sum_{n=0}^{K_0} \frac{(N - nM - 1)!}{n!(K-n)!(N-(nM+K))!}(-1)^n. \tag{65}$$

For $M = N$ we find $K_0 = 0$ and the normalization constant

$$C_{N,K,N} = a^K \frac{(N-1)!}{(K-1)!(N-K)!}, \tag{66}$$

is equal to the normalization constant $C_{N,K}$ obtained for the constant kernel in Sect. 3.1. The mean number of clusters of size $i \leq M$ conditioned on the number of clusters K is

$$\langle M_i \rangle_K = \sum_{m_i \in P'_{N,K,M}} m_i p'(m_1, \ldots m_M) = a \frac{C_{N-i,K-1,M}}{C_{N,K,M}}. \tag{67}$$

Two clusters of size i and j can form a new cluster only if $i + j \leq M$. The formation rate when there are K clusters is thus

$$f_K = \sum_{(m_i) \in P'_{N,K,M}} p'(m_1, \ldots, m_N) \left(\sum_{i=1}^{M/2} \frac{m_i(m_i-1)}{2} + \sum_{\substack{i,j=1 \\ i+j \leq M; i \neq j}}^{M} m_i m_j \right). \tag{68}$$

The formation rate can be written as a function of the coefficients $C_{N,K,M}$ as

$$f_2 = C_{N,2,M}, \tag{69}$$

and for $K > 2$

$$f_K = \frac{K(K-1)}{2} \sum_{i=1}^{\min(\frac{M}{2}, \frac{N-K+2}{2})} C_{N-2i,K-2,M} + \frac{K(K-1)}{2} \sum_{\substack{i,j=1 \\ i+j \leq M}}^{\min(M-1, N-K+1)} C_{N-i-j,K-2,M}. \tag{70}$$

The separation rate remains unchanged $s_K = (N-K)a$, and the probabilities at steady state are given by

$$\Pi_K = \frac{f_{K+1}}{s_K} \Pi_{K+1}. \tag{71}$$

We illustrate the limit case $a \to 0$ for $N = 9$, $M = 4$ (Fig. 2). When $a > 0$, all partitions are accessible, but as $a \to 0$, the steady-state configurations are dominated by the configurations with the largest possible cluster size $(4, 4, 1)$, $(4, 3, 2)$, and $(3, 3, 3)$. Applying formulas (65) and (67), we obtain the limit cluster configuration probabilities

$$\begin{aligned} p(4,4,1) &= \frac{3}{10} \\ p(4,3,2) &= \frac{6}{10} \\ p(3,3,3) &= \frac{1}{10}. \end{aligned} \tag{72}$$

These steady-state probabilities do not depend on the initial particles configurations as long as $a \neq 0$. For $a = 0$, there are three possible configurations $(4, 4, 1)$, $(4, 3, 2)$, and $(3, 3, 3)$: once equilibrium is attained, the clusters will remain unchanged. The probability to get to equilibrium depends on the configuration and the order of clustering events. When there is no limitation in the cluster formation ($M = N = 9$), a single cluster containing all particles is formed (Fig. 2, left panel). For large values of a, most clusters are very small, and the distributions are similar for $M = 4$ and $M = 9$ (Fig. 2, right panel).

The probability for two particles to be in the same cluster provides a good estimation for the cluster distribution for various values of the parameter a (Fig. 3). When a is large, most particles are contained in very small clusters and the probability $\langle P_2 \rangle$ is similar for the cases $M = 4$ and $M = 9$. When $a \to 0$, particles tend to form larger clusters. A single cluster containing all particles is formed and $\langle P_2 \rangle \to 1$ when $M = 9$, but the maximal value of $\langle P_2 \rangle$ is less than 1 when the maximal cluster size is limited. We can explicitly compute $\langle P_2 \rangle$ in the limit case $a \to 0$. For example, for $M = 4$, using Eq. (30), and summing over all possible configurations (72), we obtain

$$\begin{aligned} \langle P_2 \rangle &= p(4,4,1)P_2(4,4,1) + p(4,3,2)P_2(4,3,2) + p(3,3,3)P_2(3,3,3) \\ &= \frac{3}{10}\frac{24}{72} + \frac{6}{10}\frac{20}{72} + \frac{1}{10}\frac{18}{72} \\ &= \frac{7}{24}. \end{aligned}$$

3.3 *Example 3: Application to the Case $a_i = ai$*

We consider the case $a_i = ai$. The number of clusters of size i is asymptotically [10]

$$\langle M_i \rangle = aie^{-i\sqrt{2a/N}}. \tag{73}$$

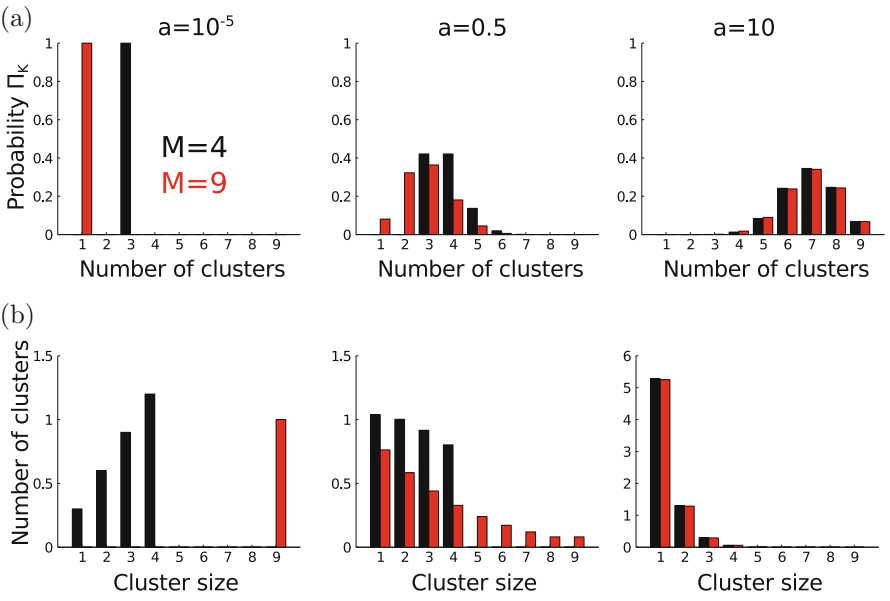


Fig. 2 (**a**) Distribution of the number of clusters Π_K for $N = 9$, when cluster sizes are limited ($M = 4$, *black*) and not limited ($M = 9$, *red*). There is a minimum of $\lceil N/M \rceil$ clusters. From *left to right*: $a = 10^{-5}$, $a = 0.5$, $a = 10$. (**b**) Mean number of clusters of each size $\langle M_n \rangle$. For $a \to 0$, for $N = 9$ and $M = 4$ the clusters organize in three different cluster configurations, while for $M = N$ a single cluster containing N particles is formed

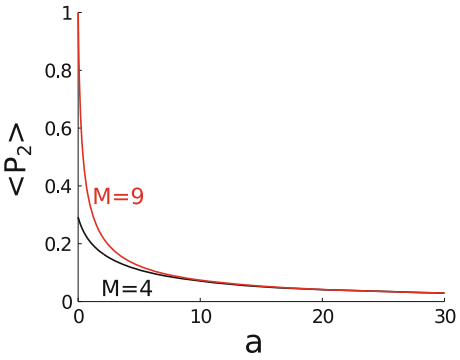


Fig. 3 Probability $\langle P_2 \rangle$ that two particles are in the same cluster. The parameters are $N = 9$ and $M = 4$ (*black*), $M = 9$ (*red*). For large values of $a \gg 1$, only small clusters are present and the steady-state distributions are similar for the cases $M = 4$ and $M = 9$. When $a \to 0$ the clusters organize in three different cluster configurations, while for $M = N$ a single cluster containing N particles is formed

Similarly to the previous examples, the number of clusters of size i, for a given distribution of K clusters, is

$$\langle M_i \rangle_{N,K} = i \frac{\binom{N-i+K-2}{N-K-i+1}}{\binom{N+K-1}{N-K}}. \tag{74}$$

The number of clusters of a given size is determined by the probability of a distribution of K clusters Π_K. It is given in the induction relation

$$\Pi_K = \frac{f_{K+1}}{s_K}\Pi_{K+1}, \tag{75}$$

where f and s are the formation and separation rates. The coagulation kernel is $C(i,j) = 1$ and the fragmentation kernel $F(i,j) = a\frac{ij}{i+j}$, and we obtain that

$$d(n) = \sum_{i=1}^{n-1} a\frac{i(n-i)}{n} = \frac{a(n^2-1)}{6}. \tag{76}$$

The separation rates are

$$s_1 = \frac{a(N^2-1)}{6} \tag{77}$$

and for $K \geq 2$

$$s_K = \frac{a}{6}\frac{1}{\binom{N+K-1}{N-K}}\frac{1}{(2K-3)!}\sum_{i=1}^{N-K+1}\frac{i(i^2-1)(N-i+K-2)!}{(N-i-K+1)!}. \tag{78}$$

In addition,

$$\begin{aligned} s_K = {} & \frac{a}{6}\frac{(2K-1)(N+K-2)}{N+K-1}((N+K-2)^2+5) \\ & -a(2N+2K-3)(2K-2) + \frac{a}{2}(N+K)^2\frac{(2K-2)(2K-1)}{2K} \\ & -\frac{a}{6}(N+K)(N+K+1)\frac{(2K-1)(2K-2)}{2K+1}. \end{aligned} \tag{79}$$

The formation rates are obtained from the number of cluster pairs that can coagulate, and are given by

$$f_K = \frac{K(K-1)}{2}. \tag{80}$$

To conclude this section, model of discrete coagulation-fragmentation processes with a finite number of particles is used to determine the steady-state probability distribution when the number of clusters is fixed. Using the partitions of the total number of particles with a given number of clusters, various statistical quantities and moments such as the cluster distributions can be computed, including also the mean number of clusters of a given size conditioned on the total number of clusters. We use two time constants to characterize the cluster dynamics: the first one is the time that two particles spend together and the second one is the time they spend separated. In the next section, we will describe specific applications in cell biology.

4 Modeling and Simulations of Telomere Coagulation-Fragmentation Process

Telomere aggregate and dissociate according to the coagulation-fragmentation process presented in Sect. 3.1. To obtain numerically any quantity of interest, we use the master equations that describe the coagulation-fragmentation process. The equation that describes the probability $P(n_1, \ldots, n_N, t)$ of having a distribution of N clusters distributed in clusters of size $n_1, \ldots, n_N$, given in Eq. (6), is t

$$\begin{aligned}\frac{d}{dt}P(n_1,\ldots,n_N,t) = &-\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} C(n_i,n_j) + \sum_{i=1}^{N}\sum_{k=1}^{n_i-1} F(k,n_i-k)\right)P(n_1,\ldots,n_N,t)\\ &+\sum_{k=1}^{N}\sum_{\substack{n_i'>0,n_j'>0\\ n_i'+n_j'=n_k}} C(n_i',n_j')P(n_1,\ldots,n_i',\ldots,n_j',\ldots,n_N,t)\\ &+\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} F(n_i,n_j)P(n_1,\ldots,n_i+n_j,\ldots,n_N,t),\end{aligned} \tag{81}$$

where $C(i,j) = k_f$ is the formation rate of a cluster of size $i+j$ from two clusters of sizes i and j, and $F(i,j) = k_b$ is the rate of dissociation of a cluster of size $i+j$ into two clusters of size i and j.

Because the time distribution of the telomere to a small target is exponential, the encounter rate of telomeres at the nuclear periphery can be characterized by a single parameter (the arrival rate or equivalently by an effective diffusion constant). Even though telomere motion involves complex polymer chains accounting for the physical chromosomal chain, any encounter is a rare event, and its rate is Poissonian. Consequently, to model clustering, we use this property to approximate the arrival time of a chromosome to a small cluster by the Poissonian dynamics, as long as the chromosome length does not restrict the motion of the telomere on the nuclear surface. Two telomeres encounter at a Poissonian rate k_f.

Polymer simulations (Fig. 4b) confirm that the arrival time of a telomere to a cluster can be simulated using a Poissonian distribution approach. In that case, it is enough to study the dynamics of 32 stochastic particles (Fig. 4c). Thus, using a molecular dynamics simulation of two Brownian particles on the surface of a sphere [6], Brownian simulations of particle located on the two-dimensional sphere except for a region of the size of the nucleolus (see earlier discussion) lead to an approximation for the forward rate of $k_f \approx 1.9 \times 10^{-3}\,\mathrm{s}^{-1}$, where the encounter disk is of radius $\delta = 0.015\,\mu\mathrm{m}$ and the effective diffusion constant is $D = 0.005\,\mu\mathrm{m}^2/\mathrm{s}$ [5].

When a telomere aggregates to a cluster, it only slightly varies in size. Indeed, in the complex environment of the nuclear surface, the diffusion constant varies with the log of the radius of the effective diffusing particle. Thus any changes in the radius

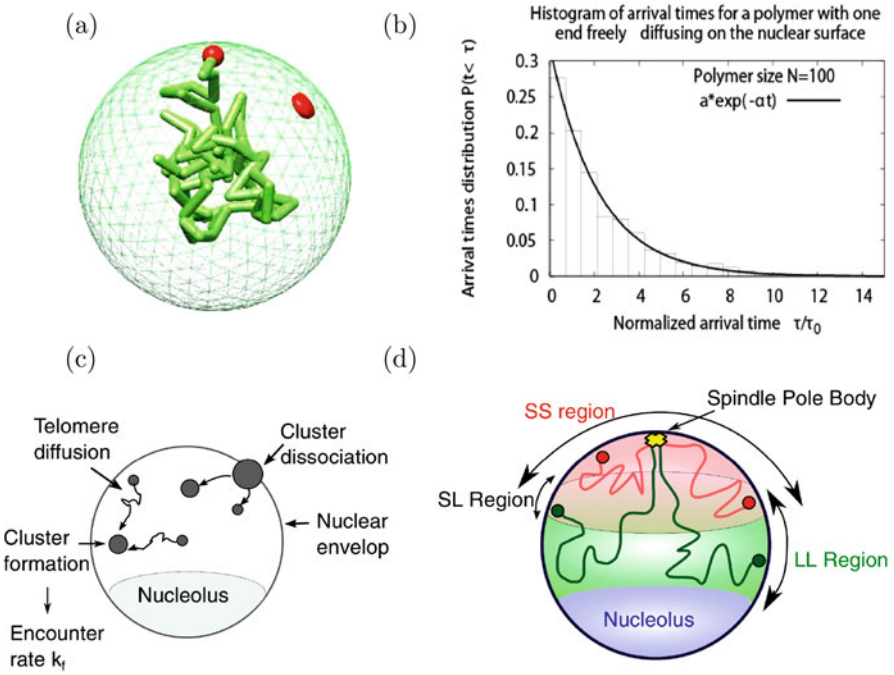


Fig. 4 Computational model of telomere cluster formation. (**a**) Snapshot from a Brownian dynamics simulation of a polymer with one end anchored on the nuclear surface. The polymer is composed of 100 monomers with average distance between monomers of $l_0 = 50\,\text{nm}$, and the nucleus is a reflecting sphere of size $R = 250\,\text{nm}$. (**b**) Histogram of the arrival times for a polymer of 100 monomers freely diffusing in the nucleus and one end constrained to diffusion on the surface. A fit of the form $f(t) = a\exp(-bt)$ gives $a = 1.014$ and $b = 0.76$. (**c**) The diffusion-aggregation-dissociation model of telomere organization. Telomeres are simplified as Brownian particles diffusing on the nuclear surface that can meet and form clusters, and clusters of n telomeres split at a rate $(n-1)k_b$. The coarse-grained association rate k_f is taken as the average of the cluster meeting times. (**d**) Influence of long and short chromosome arms on clustering. Decomposition of the nucleus in subdomains with telomeres from short and long chromosome arms. Both types of telomere can interact in a common region

will result only in a small change in the diffusion coefficient. We neglected any possible changes in the scattering cross section and motility, which could modify the forward binding rate [14]. Thus the encounter rate between clusters or telomeres will be approximated by a constant independent of the size.

In the Gillespie's algorithm, the transition rate constants between different cluster configurations are given as follows: for a distribution $(n_1, \ldots, n_K)$ of clusters, the transition probabilities to the neighboring states depend on two events: either two clusters (n_i, n_j) associate to form a new cluster of size $n_i' = n_i + n_j$ with an association rate k_f or a cluster of size n dissociates into two, with a rate $(n-1)k_b$ that depends on the number of bonds. The size of the resulting dissociated clusters is uniformly distributed in the interval $[1, n-1]$. Since there are $\frac{K(K-1)}{2}$ pairs,

the association rate equals $\frac{K(K-1)}{2}k_f$, and the total fragmentation rate is the sum over all dissociation rates $\sum_j (n_j - 1)k_b = (N - K)k_b$. The total transition rate from the state $(n_1, \ldots, n_K)$ to any of the possible association and dissociation events is $a_0(n_1, \ldots, n_K) = \sum a_i = \frac{K(K-1)}{2}k_f + (N - K)k_b$. Each iteration step of the algorithm uses the classical Poissonian random transition time $\tau = -\frac{\log r_1}{a_0}$, where r_1 is a uniform random variable in $[0, 1]$ and each reaction event i has a probability $\frac{a_i}{a_0}$ to occur, and the chosen reaction i is sorted out using the criteria $\sum_{j=1}^{i-1} \frac{a_j}{a_0} < u \leq \sum_{j=1}^{i} \frac{a_j}{a_0}$ where u is uniformly distributed in $[0, 1]$.

4.1 *Influence of the Chromosome Arm Length on the Clustering Dynamics*

Because chromosome arms with a length below 300 kb are mainly located in a small region near the spindle pole body (SPB) [31], while telomeres of longer chromosome arms exhibit motion near the nucleolus, we decided to integrate these constraints into the telomere dynamics (Fig. 4d). We distribute telomeres into two classes based on the length of the chromosome arm [31] and restricted 12 telomeres to a small region account for short–short interactions (SS) around the SPB (1/3 of the surface) and the other 20 are free to diffuse in a larger region where only long telomeres can interact (LL), which excludes both the nucleolus and a small cap around the SPB (SL is 2/3 of the nucleus surface).

In the common region SL, both types of telomeres can meet to form mixed clusters. There are three possible classes of telomere clusters: clusters containing telomeres from long chromosome arms only (long), from short chromosome arms only (short), or from long and short chromosome arms (mixed), leading to six forward rates, accounting for the long–long, short–short, long–short, long–mixed, short–mixed, and mixed–mixed interactions.

In addition, for two telomeres from the pool of long chromosome arms, the recurrence time is $T_R = 442$ s ($n = 1000$), shorter than the forward time $k_f^{-1}(L, L) \approx 500s$. Thus, the interaction of telomeres from short chromosome arms with a cluster made of long ones will contribute to the confinement of the cluster to a smaller region of the nuclear periphery, which will consequently decrease the mean time for two telomeres to meet again. The mean time to separation T_S was similar for telomeres from short–short, short–long, and long–long chromosome arms ($\approx$21 s, versus 31 s for the dissociation time between two telomeres, $n = 1000$), reflecting that clusters contain the same number of telomeres independently of their composition.

Finally, the equilibrium probability to find a given telomere in a visible cluster (containing more than 2) was $Pr(S, S) = 0.06$, $Pr(L, L) = 0.045$, and $Pr(L, S) = 0.04$ (for short–short, long–long, and long–short arm interactions), confirming that the encounter rate for small telomeres is higher than for long ones, due to the smaller space they can explore. Our results are mainly consistent with [31],

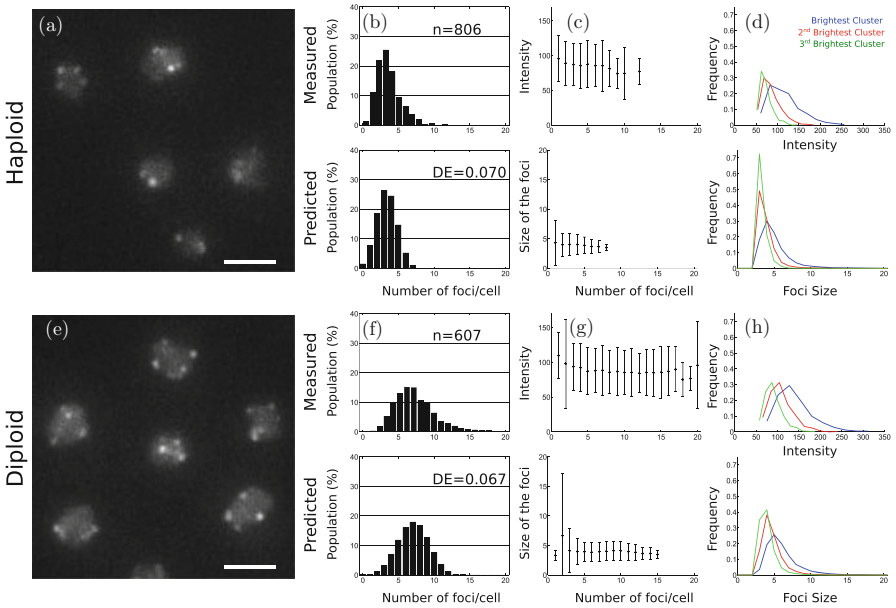


Fig. 5 Comparison of experimental and simulation results of telomeres clustering in yeast. (**a**, **e**) Live cell imaging of telomere clusters. Representative fluorescence image of the telomere-associated protein Rap1 tagged with GFP (scale bar, 2 μm) in haploid (**a**) and diploid cells (**b**). (**b**, **f**) Histogram of the number of clusters per cell. (**c**, **g**) Mean ± s.d. of the intensity distributions of the clusters in live cells and distribution of the cluster size in the Brownian simulations. In the haploid cells, clusters are made of four telomeres, with a small dispersion that does not depend on the cluster number. (**d**, **h**) Fluorescence intensity (experiments) and sizes defined as the number of telomeres per cluster (simulations) for the three brightest clusters. The frequency of occurrence (*y*-axis) of a given cluster size is plotted as a function of the intensity of a cluster (*x*-axis), proportional to the telomere number

where the probabilities for two telomeres belong to the same focus are determined experimentally to be mostly in the range 0.04–0.09. The differences between these experimental data and our simulations might be due to specific interactions between telomere pairs, which we did not take into account. Indeed, contacts between telomeres on opposite chromatid arms of equal length are favored [27].

The aggregation-dissociation model for telomere organization was used to extract in vivo parameters by comparing stochastic simulations with live cell imaging data (Fig. 5a). The dissociation rate k_b is estimated by comparing the experimental and simulation histograms for the number of clusters containing more than two telomeres (Fig. 5b). Histograms similarity was evaluated using the Kolmogorov–Smirnov (KS) score, here defined as the maximum of the absolute difference of the experimental and simulated cumulative distribution function for the number of clusters. The optimal value of the KS score was 0.11 obtained for $k_b = 2.4 \times 10^{-2}\,\mathrm{s}^{-1}$.

However, a higher variance in the histogram of the experimental number of clusters. To account for this variation, we introduced fluctuations in the value of

the dissociation rate k_b of each cell. We generated random values of k_b following a Gaussian distribution and we found that for $k_b = 2.310^{-2} \pm 1.3 \times 10^{-2}\,\mathrm{s}^{-1}$, which corresponds to $a = k_b/k_f = 12 \pm 7$, we obtain an optimal fit for the distribution of the number of clusters. Simulations show an excellent adequacy to the experimental cluster distribution (Fig. 5b), size (Fig. 5c), and size distribution (Fig. 5d), with a KS score of 0.07.

We observed an average of three detectable clusters per cell, and very few cells with more than eight clusters. Interestingly, in the simulations, 9.9 (8.2) telomeres are isolated (in pairs). In addition, the number of telomeres per cluster obtained in our simulations reflects very well the cluster intensity obtained experimentally: in both simulated and experimental data, we found that the average cluster intensity does not vary with the number of clusters per cell (Fig. 5c). Because there are 32 telomeres and that the intensity is an increasing function of the number of telomeres, we conclude that there are in average no more than four telomeres per cluster. A better precision about the cluster distribution is obtained by plotting the distribution of the first three brightest clusters for both experimental and simulated data (Fig. 5d): in both cases the three brightest clusters contain four telomeres.

The robustness of the aggregation-dissociation model is tested for the organization of telomeres in diploid cells where the nuclear volume (nucleus radius = $1.25\,\mu\mathrm{m}$) and the number of telomeres are doubled. These changes in the cell geometry affect the forward rate, which we recomputed from Brownian simulations, and we now found for the association rate $k_f = 1.1 \times 10^{-3}\,\mathrm{s}^{-1}$. Considering that the backward rate is unchanged and taking the value found in the normal case, we obtained for the new equilibrium constant the value $a = 21 \pm 12$ (compared to 12 ± 7 for the haploid).

Telomere foci in diploid cells are shown in Fig. 5e, and the number of telomere foci obtained by simulation is similar to the number measured in live cells. They have in average six clusters containing three to six telomeres per cell (Fig. 5f, g). The light intensity and the telomeres distribution of measured and simulated telomeres per cluster were very similar (Fig. 5f–h). Interestingly the median cluster size is 4 in both haploid and diploid cells, i.e., there are four telomeres per cluster, suggesting that the number of telomeres per cell does not influence the number of telomeres per cluster. Furthermore, according to the simulations, in diploid cells, telomeres cluster in 5–9 foci containing 3–6 telomeres, while 18.7 telomeres are single and 16.4 are in pairs. The matching between experimental data and numerical simulations confirms the robustness of the model to parameter changes, while the physical properties of the telomeres and the cluster dissociation rate were maintained fixed.

5 Modeling Capsid Formation as an Aggregation Process

In this second part, we study here the kinetics of cluster formation starting with the arrival and fusion of elementary particles at a nucleation site. Particles involved in the cluster formation are organized in aggregates. The aggregates increase the

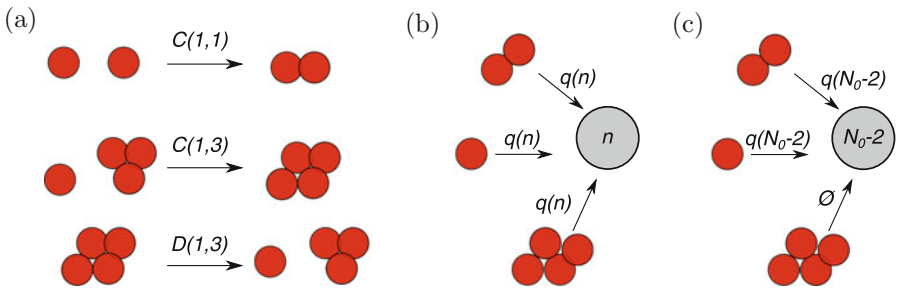


Fig. 6 Schematics of clustering for a finite number of particles. (**a**) Model used for telomere clusters coagulation-fragmentation. (**b**, **c**) Model used for capsid formation

cluster size by fusion to the particles. This is an elementary model of capsid viral assembly, where the density distribution of aggregates is at steady state. To maintain this distribution, the protein production must be much larger than the aggregates needed to form a single capsid.

In that model, a cluster can accept a maximum of N_0 particles, and is complete when exactly N_0 particles have arrived. The cluster is formed upon the arrival of aggregates of various size. When a cluster has reached a size n, it can accept aggregates of size less than $N_0 - n$ (Fig. 6c). Each aggregate binds to a cluster with a Poissonian arrival rates λ, independent of the aggregate size. Aggregates participating in the cluster formation are already formed and are at steady state. Therefore the number n_k of aggregates containing k particles is constant. The total number of particles N_T is distributed among the aggregates, therefore

$$N_T = \sum_{k=1}^{N_0} k n_k. \tag{82}$$

We assume that the number of aggregates of size k is distributed exponentially and given by

$$n_{k+1} = p n_k, k \geq 0 \tag{83}$$

where the parameter $0 \leq p \leq 1$.We present here the models of nucleation using a mean-field approximation and a stochastic jump process.

5.1 Mean-Field Approximation

We now derive an equation for the cluster size $n(t)$ at time t. The cluster growth rate depends on the arrival of an aggregate of size k and on the probability q of finding a free site at the cluster. We neglected here the geometrical organization

of an aggregate and consider that upon fusion, it fills empty slots in the cluster. We do account here for the geometrical organization in facet of aggregates which participate to the structure of viral capsids. Thus, the probability q does not depend on the geometry or positions of the aggregates already present in the cluster but only on their number. We chose the linear relation

$$q(n(t)) = 1 - \frac{n(t)}{N_0}. \tag{84}$$

In addition, we neglected any changes of the arrival rate due to the size of the aggregate that can affect the diffusion coefficient. To conclude, a nucleation site is formed when it is entirely filled by aggregates. The cluster growth is due to the arrival of aggregate of size k and the rate is $\lambda k n_k$. The cluster total growth rate is the product of the probability to find an available site times the sum of the arrival rate of any aggregate. The average size $n(t)$ satisfies the equation

$$\dot{n}(t) = \lambda \left(1 - \frac{n(t)}{N_0}\right) \left(\sum_{k=1}^{N_0 - n(t)} k n_k\right), \tag{85}$$

which reduces to

$$\dot{n}(t) = A(N_0 - n(t)) \left[1 - p^{N_0 - n(t)}(1 + (N_0 - n(t))(1-p))\right], \tag{86}$$

with initial condition $n(0) = 0$ and

$$A = \frac{\lambda N_T}{N_0 \left[1 - p^{N_0}(1 + N_0(1-p))\right]}. \tag{87}$$

For $0 < p < 1$, although Eq. (86) cannot be integrated analytically, we obtain the short and long time asymptotic in the limit of N_0 large. For short time, the size for the growing cluster is

$$n(t) \approx \lambda N_T t, \text{ for } t \ll 1, \tag{88}$$

which is independent of p. For large t and small p, the first order expansion is

$$n(t) \approx N_0 - \frac{N_0}{\lambda N_T (p - 1 - \log p)} \frac{1}{t} \text{ for } t \gg 1. \tag{89}$$

In the limit case $p = 1$, Eq. (86) changes its nature and for large N_0, it reduces to

$$n(t) = N_0 - \frac{N_0}{\sqrt{1 + 2\lambda \frac{N_T}{N_0} t}} \text{ for } t \gg 1. \tag{90}$$

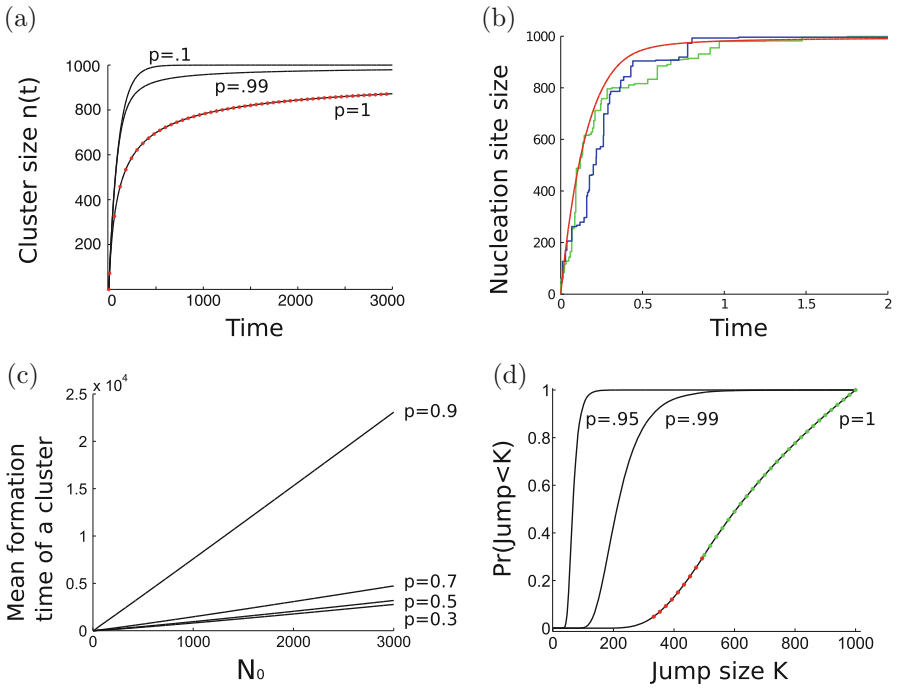


Fig. 7 (**a**) Kinetics of the cluster growth in the mean-field approximation. Various kinetics profile [solution of Eq. (86)] for $N_T = 1000$, $N_0 = 1000$, $\lambda = 10^{-2}$, and $p = 0.1, 0.99, 1$. (**b**) Comparison of the kinetics of formation with the deterministic and the stochastic models. Parameters are $N_0 = 1000$, $p = 0.96$, $\lambda = 0.001\,\mathrm{s}^{-1}$. (**c**) Mean time to form a function as a function of N_0 for various values of p, with $\lambda = 0.001\,\mathrm{s}^{-1}$ and $N_T = 1000$. (**d**) Cumulative distribution of the maximal jump size $P_{N_0,K}$, for $N_0 = 1000$ and $p = 0.95, 0.97, 0.99, 1$. The probability $P_{N_0,K}$ for $p = 1$ is compared with the approximation Eqs. (104) (*green*) and (105) (*red*)

When there are single monomers only ($p = 0$) Eq. (85) describes the classical kinetics of arrival and the solution to Eq. (86), which reduces to a single exponential $n(t) = N_0(1 - e^{-\lambda \frac{N_T}{N_0} t})$. We plotted in Fig. 7a the kinetics of the cluster formation. Interestingly, the cluster is formed more quickly for lower values of the parameter p.

5.2 *A Stochastic Dynamics for the Cluster Formation*

Due to the discrete arrival of aggregates to the nucleation site, the cluster size increases by random jumps that we shall describe now using a stochastic jump process. When an aggregate arrives in the time interval $(t, t + \Delta t)$, the cluster of size $n(t)$ at time t increases with a probability $\mu(n(t))dt$ that depends on its size at time t, thus

$$n(t+\Delta t) = \begin{cases} n(t) \text{ w.p. } 1-\mu(n(t))\Delta t \\ n(t)+J(n(t)) \text{ w.p. } \mu(n(t))\Delta t, \end{cases}$$

where $J(n(t))$ is the size of a random jump, characterized by its conditional transition distribution

$$\begin{aligned} Pr\{J(n(t)) = m-n | n(t) = n\} &= w(m-n|n) \\ &= \frac{(1-p)p^{m-n}}{p(1-p^{N_0-n})} \end{aligned}$$

and $w(m-n|n)$ is transition probability from n to m, which we normalized by summing over all aggregate sizes that can fuse with the cluster. To determine the arrival rate of an aggregate, we start with a cluster containing n particles. The arrival rate of an aggregate is given by the jump rate $\mu(n)$, which is equal to the arrival rate λ of an aggregate of particles (or a single particle), multiplied by the number of aggregates smaller than $N_0 - n$, so they can enter in the nucleation site, multiplied by the probability of finding a free site (proportional to $1 - \frac{n}{N_0}$). The jump rate is thus

$$\begin{aligned} \mu(n) &= \lambda \left(1 - \frac{n}{N_0}\right) \sum_{k=1}^{N_0-n} n_k \\ &= a(N_0 - n)(1 - p^{N_0-n}), \end{aligned} \tag{91}$$

where $a = \lambda \frac{N_T}{N_0} \frac{1-p}{1-p^{N_0}(1+N_0(1-p))}$. The probability density function satisfies the master equation

$$\begin{aligned} p(m, t+\Delta t) &= (1-\mu(m)\Delta t)p(m,t) \\ &\quad + \sum_{n=1}^{m-1} w(m-n|n)p(n,t)\mu(n)\Delta t, \end{aligned}$$

which tends in the limit $\Delta t \to 0$, to the discrete forward Fokker–Planck equation

$$\begin{aligned} \frac{\partial p(m,t)}{\partial t} &= L_m p(m,t) = -\mu(m)p(m,t) \\ &\quad + \sum_{n=1}^{m-1} \mu(n)w(m-n|n)p(n,t), \end{aligned} \tag{92}$$

where L_m is the forward Kolmogorov operator.

5.3 The Mean Time to Cluster Formation

The time to form the cluster is the mean first passage time $\langle \tau(n) \rangle$ of the cluster size to its maximum N_0. By definition,

$$\tau(n) = \inf \{t > 0; n(t) \geq N_0 | n(0) = n\} \tag{93}$$

and the MFPT is solution of the backward equation [29] with absorbing boundary condition at N_0

$$\begin{cases} L_n^* \langle \tau(n) \rangle = -1 \\ \langle \tau(N_0) \rangle = 0, \end{cases} \tag{94}$$

where the operator L_n^* is the adjoint of L_m. The MFPT is obtained by solving the system of equations for $0 \leq n \leq N_0 - 1$,

$$-1 = -\mu(n) \langle \tau(n) \rangle + \sum_{m=n+1}^{N_0} \langle \tau(m) \rangle \mu(n) w(m-n|n). \tag{95}$$

The mean time of a cluster formation is then

$$\langle \tau(0) \rangle = \frac{N_0(1 - p^{N_0}(1 + N_0(1-p)))}{\lambda N_T} \times \left[\frac{1}{N_0(1-p)(1-p^{N_0})} + \sum_{i=1}^{N_0 - 1} \frac{1}{i(1-p^{i+1})(1-p^i)} \right], \tag{96}$$

which depends on the total number of particles N_T, the maximal size N_0, the parameter p that describes the size distribution of aggregates, and λ the arrival rate of an aggregate to the nucleation site.

For small p and large N_0, we obtain the approximation

$$\langle \tau(0) \rangle = \frac{N_0}{\lambda N_T} (\log N_0 + \gamma) + o(p). \tag{97}$$

For a large nucleation site N_0, in the limit $p \to 1$, the mean time remains finite and Eq. (96) becomes

$$\langle \tau(0) \rangle (N_0, p, \lambda, N_T) = \frac{N_0^2}{\lambda N_T} \left(\frac{\pi^2}{6} - 1 \right) \text{ for } p = 1. \tag{98}$$

The mean formation time does increase drastically as p tends to 1 (Fig. 7c). Indeed, a cluster starts growing very rapidly when large aggregates arrive, however, the

growth is reduced later on because the number of admissible aggregates (smaller than the number of available sites) is small. Admissible aggregates represent only a small fraction of the total number of aggregates.

5.4 Composition of a Cluster

We now characterize the cluster assembly by studying the size distribution of aggregates that have arrived to the nucleation site. We shall derive also the size of the largest aggregate that contributes to the cluster formation. To evaluate the various sizes of aggregates that bind to the cluster, we consider the ensemble of aggregates. During the sequential steps of aggregation, the number of particles C_n at the nth-step, with $C_0 = 0$, follows the equation:

$$C_{i+1} = C_i + z_{i+1}, \tag{99}$$

when z_{i+1} is the size of the aggregate that binds at step $i + 1$. The cluster contains a maximum of N_0 particles. The size of aggregate z_{i+1} that binds to the cluster can take values in $(1, \ldots, N_0 - C_i)$ and thus the probability that the $i + 1$th aggregate is of size k when there are $N_0 - C_i$ free sites is

$$P_{N_0-C_i}(z_{i+1} = k) = \frac{(1-p)p^{k-1}}{1-p^{N_0-C_i}}. \tag{100}$$

Thus, the joint probability that the cluster assembles with the following order of arrival $(k_1, \ldots, k_n)$ is the product of the conditional probabilities (100)

$$P(z_1 = k_1, \ldots, z_n = k_n) = \prod_{i=1}^{n} P_{N_0-\sum_{j=1}^{i-1} k_j}(z_i = k_i),$$

with the condition $\sum_{i=1}^{n} k_i = N_0$.

5.5 The Largest Aggregate Merging to the Cluster

The probability that the largest aggregate $z_{\max}$ is less than K during the cluster assembly is

$$P_{N_0,K} = \sum_{\substack{\{(k_1,\ldots k_n);\sum k_i = N_0\} \\ \text{and } k_1,\ldots,k_n \le K}} P(z_1 = k_1, \ldots, z_n = k_n). \tag{101}$$

To obtain an approximation of $P_{N_0,K}$ for large N_0, we sum over the first jump size, which leads to the induction formula

$$P_{N_0,K} = \sum_{k_1=1}^{K} P_{N_0}(z_1 = k_1) P_{N_0-k_1,K}, \tag{102}$$

where $P_{n,k} = 1$ for $n \leq k$. When the parameter $p = 1$, formula (102) reduces to

$$P_{N_0,K} = \sum_{k_1=1}^{K} \frac{1}{N_0} P_{N_0-k_1,K}. \tag{103}$$

The induction formula (103) can be solved by $P_{N,K} = f(\frac{K}{N})$, where f satisfies $f'(x) = \dfrac{f(\frac{x}{1-x})}{x}$ with $x = \frac{K}{N}$. The function f is solution of a nth order linear differential equation on each interval $[\frac{1}{n+1}, \frac{1}{n}]$ for $n \geq 1$, which we solved on the intervals $(\frac{1}{3}, \frac{1}{2})$ and $(\frac{1}{2}, 1)$. However, there is no simple formula on other intervals. Finally, the probability for the size of the largest aggregate to be less than K after the cluster is filled is given for large N_0 by

$$P_{N_0,K} = 1 + \log \frac{K}{N_0} \text{ for } N_0/2 \leq K \leq N_0 \tag{104}$$

and

$$P_{N_0,K} = 1 + \log \frac{K}{N_0} + \mathrm{Li}_2\left(\frac{K}{N_0}\right) + \frac{1}{2}\log^2 \frac{K}{N_0} + \log\left(\frac{K}{N_0}\right) + 1 \text{ for } N_0/3 \leq K \leq N_0/2 \tag{105}$$

where $\mathrm{Li}_2(x) = \sum_{k=1}^{\infty} \frac{x^k}{k^2}$. The probability $P_{N_0,K}$ is well approximated by the function f that we constructed inductively on intervals $(\frac{1}{3}, \frac{1}{2})$ and $(\frac{1}{2}, 1)$ (Fig. 7d). The construction of the function f reflects that the number of possible jumps of maximal size is limited: indeed, once an aggregate of size larger than $N_0/2$ has arrived, the size of all other aggregates can only be smaller than $N_0/2$, leading to the initial interval $(\frac{1}{2}, 1)$. Similarly, after an aggregate of size between $N_0/3$ and $N_0/2$ has arrived, other possible aggregates have a size smaller than $N_0/3$. This constraint leads to the second interval $(\frac{1}{3}, \frac{1}{2})$. We obtain by induction the division in intervals $(\frac{1}{n+1}, \frac{1}{n})$.

5.6 *Gag Protein Aggregation in Potential Wells*

Recent super-resolution data have revealed that GAG proteins of the HIV virus can aggregate in specific microdomains [12]. Interestingly, the proteins aggregate in small regions characterized by a physical potential well (Fig. 8), discovered in [14]. Indeed the motion of aggregates on the membrane surface is influenced by a diffusion coefficient D and a field of force $F(X, t)$, following the overdamped Langevin model equation

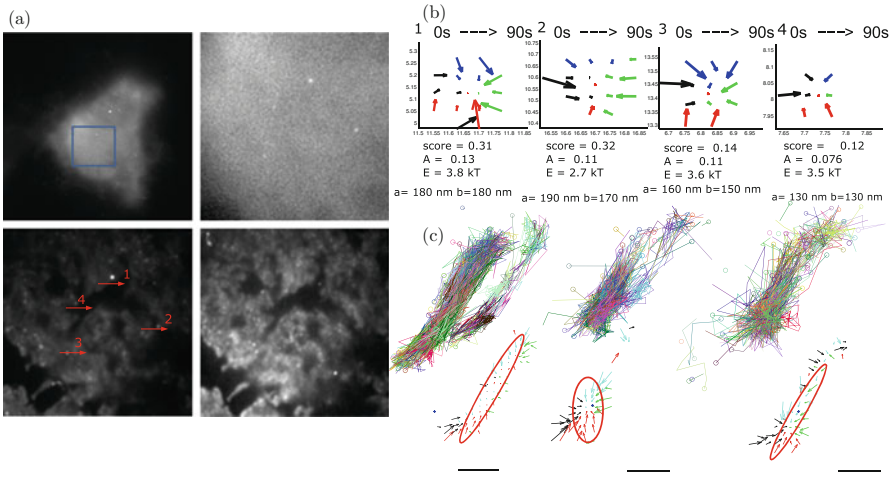


Fig. 8 **(a)** Area of aggregation (*right*) correspond to potential wells, characterized by converging *arrows*. **(b)** Four examples are shown (*left*) with associated parameters of the ellipses: long a and short b abscise. **(c)** Three other potential wells represented with the high density of GAG trajectories, scale bar 500 nm (data given by the courtesy of S. Manley)

$$\dot{X} = \frac{F(X(t), t)}{\gamma} + \sqrt{2D}\,\dot{W}, \tag{106}$$

where W is a Gaussian white noise and γ is the dynamical viscosity [28]. The source of the noise is the thermal agitation of the ambient lipid and membrane molecules. However, at low resolution, the motion is described by an effective stochastic equation [15, 19]

$$\dot{X} = a(X)dt + \sqrt{2}B(X)\dot{W}, \tag{107}$$

where $a(X)$ is the drift field and $B(X)$ the diffusion matrix. The effective diffusion tensor is given by $D(X) = \frac{1}{2}B(X)B^T(X)$ ($.^T$ denotes the transposition) [28, 30]. The observed effective diffusion tensor is not necessarily isotropic and can be state-dependent, whereas the friction coefficient γ in (106) remains constant and the microscopic diffusion coefficient (or tensor) may remain isotropic.

The drift field $a(\boldsymbol{x})$ in Eq. (107) represents a force that acts on the diffusing particle, regardless of the existence or not of a potential well [13]. In the case where $\boldsymbol{D}(\boldsymbol{x})$ is locally constant and the coarse-grained drift field $\boldsymbol{b}(\boldsymbol{x})$ is a gradient of a potential

$$a(\boldsymbol{x}) = -\nabla U(\boldsymbol{x}), \tag{108}$$

then the density of particles represents locally the Boltzmann density $e^{-U(\boldsymbol{x})/D}$ [13]. The force field can form potential wells, generically approximated locally as a

paraboloid with an elliptic base. It remains a difficult question to extract the axis, the center, and the boundary of the elliptic base of the well. Once they are known, within the analytical representation $U(\boldsymbol{x}) = A\left((\frac{x}{r_x})^2 + (\frac{y}{r_y})^2\right) + O(x,y)^2$, the constants A, r_x, r_y are three parameters to be determined.

GAG proteins show free and confined motions. The density of particles is quite heterogeneous, with many small dense regions and a few very dense regions (Fig. 8). The stochastic analysis and diffusion map (Fig. 8) reveals a mean diffusion coefficient is $D = 0.7\,\mu\text{m}^2/\text{s}$ almost uniform. Several potential wells could be detected with an elliptical base with radius 170–200 nm. One with depth $A = 0.78\,\mu\text{m}^2/\text{s}$ with a score of 0.20, confirming that these wells are robust [19]. The energy of the potential well is in the range 1.7–4 kT.

Interestingly, the wells evolve in time (Fig. 9) and can disappear rapidly (in less than 5 min) and the energy decreases gradually in time. This analysis used a moving

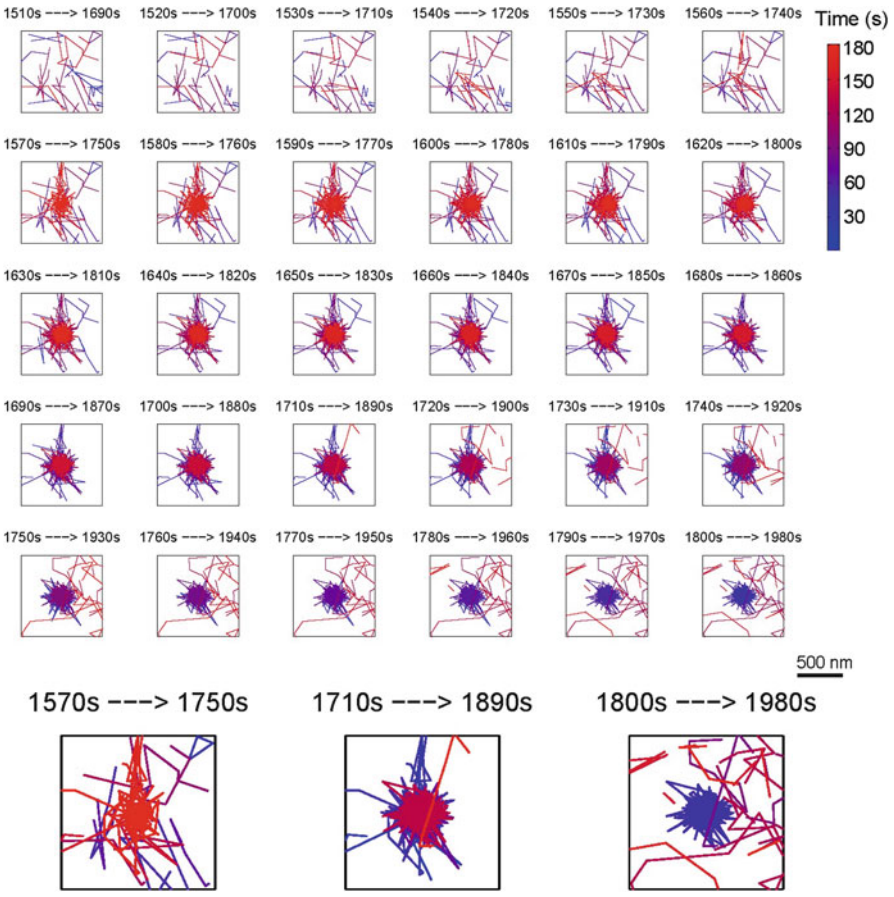


Fig. 9 (*Upper*) Time dependent aggregation, separated in time windows. *Lower* (**a**, **b**) transient potential wells corresponding to panel **(a)**. For the last panel, some GAG trajectories are not attracted by potential wells (data given by the courtesy of S. Manley)

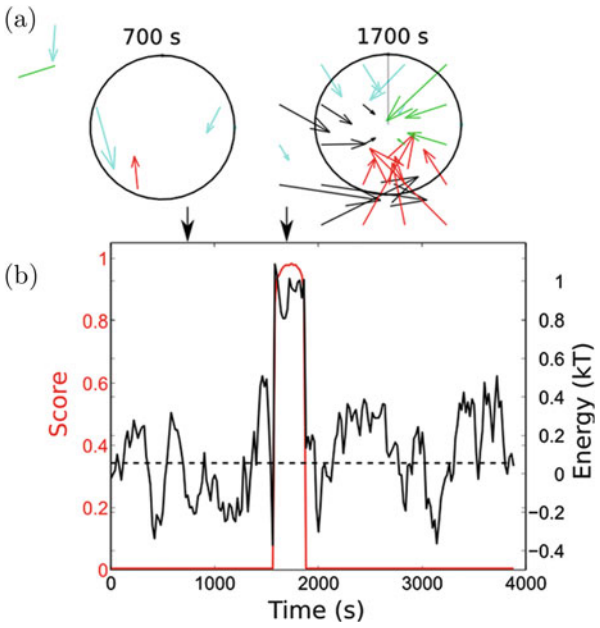


Fig. 9 (continued)

window, which smooths out fluctuations. To observe the evolution of the trajectories in a small region in the proximity of the potential well, we plotted windows of 180 s of recording (Fig. 9). For each panel, the trajectories were recorded in the time interval $(t, t + 180\,\text{s})$. The next panel represents trajectories taken 10 s afterwards, in the time interval $(t + 10\,\text{s}, t + 190\,\text{s})$. To represent the evolution of trajectories through time, in each window, the trajectories are colored during the first seconds in blue, and trajectories near 180 s in red. The most recent trajectories are overlaid on the first trajectories.

In the first seconds, trajectories appear unorganized (Fig. 9, top row). Confined trajectories appear in only 10 s at 1750 s (Panel 1570–1750 s). This confinement lasts for 140 s: after time 1890 s (Panel 1710–1890 s), the new trajectories (red) that pass over the former confinement region are diffusive and not attracted to any point. To conclude, the potential well lasted for 140 s between 1750 and 180 s. Moreover the confinement region is expanding through time. The radius of the potential well changed from approximately 200 nm at the beginning at 1750 s (Panel 1570–1750 s) to a radius of 250 nm at 1890 s. To measure the changes in energy of the well through time, the energy of the well in each window of 180 s is shown in Fig. 9 lower panel. During the time interval (1750–1890 s), which corresponds to the period of confinement, the proximity of the measured drift map with a parabolic expression is in very good agreement in the confinement in the time period (1750–1890 s). This agreement confirms the presence of interaction forces acting on the Gag proteins. Finally, the present analysis confirms that aggregate formation occurs in geometrical confined structures that are transient in time.

5.7 Conclusion

We presented here several analytical formula based on aggregation-fragmentation with a finite number of particles. These formulas can be used to extract parameters such as rate constants from experimental data. The general framework is also used to derive the extreme statistics about the time formation of a cluster or the time two particles spend in the same cluster.

We also discussed here two important applications about telomere clustering in yeast [14, 20] and capsid formation [17]. The geometrical organization of a cluster formation from small aggregates remains difficult to account for into modeling. Future directions should be concerned with accounting for the random geometry of aggregates and their insertion in a cluster. In the last subsection, we reviewed experimental evidences that capsid assembly might use the membrane local curvature, but the exact mechanism remains open.

Acknowledgements This research was supported by a Marie-Curie grant. We thank S. Manley for discussions and sharing with us the GAG super-resolution data. We also thank the hospitality of the Newton Institute in Cambridge during the year 2016.

References

1. M. Abramowitz, I.A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Reprint of the 1972 edn. (Dover, New York, 1992)
2. D.J. Aldous, Deterministic and stochastic models for coalescence (aggregation, coagulation): review of the mean-field theory for probabilists. Bernoulli **5**, 3–48 (1999)
3. G.E. Andrews, *The Theory of Partitions. Encyclopedia of Mathematics and Its Applications*, vol. 2 (Addison-Wesley, Reading, MA, 1976)
4. R. Becker, W. Döring, Kinetische Behandlung der Keimbildung in übersättigten Dämpfen. Ann. Phys. **24** 719–752 (1935)
5. K. Bystricky, P. Heun, L. Gehlen, J. Langowski, S.M. Gasser, Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. Proc. Natl. Acad. Sci. U. S. A. **101**(47), 16495–16500 (2004)
6. T. Carlsson, T. Ekholm, C. Elvingson, Algorithm for generating a Brownian motion on a sphere. J. Phys. A Math. Theor. **43**(50), 505001 (2010)
7. S. Chandrasekar, Stochastic problems in physics and astrophysics. Rev. Mod. Phys. **15**, 1–89 (1943)
8. J.F. Collet, Some modelling issues in the theory of fragmentation-coagulation systems. Commun. Math. Sci. **1**, 35–54 (2004)
9. C.R. Doering, D. Ben-Avraham, Interparticle distribution functions and rate equations for diffusion-limited reactions. Phys. Rev. A **38**, 3035 (1988)
10. R. Durrett, B.L. Granovsky, S. Gueron, The equilibrium behavior of reversible coagulation-fragmentation processes. J. Theor. Probab. **12**, 447–474 (1999)
11. S. Gueron, The steady-state distributions of coagulation-fragmentation processes. J. Math. Biol. **37**, 1–27 (1998)
12. J. Gunzenhäuser, R. Wyss, S. Manley, A quantitative approach to evaluate the impact of fluorescent labeling on membrane-bound HIV-Gag assembly by titration of unlabeled proteins. PLoS One **9**(12), e115095 (2014)

13. D. Holcman, N. Hoze, Z. Schuss, Analysis and interpretation of superresolution single-particle trajectories. Biophys. J. **109**, 1761–1771 (2015)
14. N. Hoze, D. Holcman, Coagulation–fragmentation for a finite number of particles and application to telomere clustering in the yeast nucleus. Phys. Lett. A **376**, 845–849 (2012)
15. N. Hoze, D. Holcman, Residence times of receptors in dendritic spines analyzed by stochastic simulations in empirical domains. Biophys. J. **107**, 3008–3017 (2014)
16. N. Hoze, D. Holcman, Modeling capsid kinetics assembly from the steady state distribution of multi-sizes aggregates. Phys. Lett. A **378**, 531–534 (2014)
17. N. Hoze, D. Holcman, Kinetics of aggregation with a finite number of particles and application to viral capsid assembly. J. Math. Biol. **70**, 1685–1705 (2015)
18. N. Hoze, D. Holcman, Stochastic coagulation-fragmentation processes with a finite number of particles. Ann. Appl. Probab. (2016, accepted)
19. N. Hoze, D. Nair, E. Hosy, C. Sieben, S. Manley, A. Herrmann, J.B. Sibarita, D. Choquet, D. Holcman, Heterogeneity of receptor trafficking and molecular interactions revealed by superresolution analysis of live cell imaging. Proc. Natl. Acad. Sci. U. S. A. **109**, 17052–17057 (2012)
20. N. Hoze, M. Ruault, C. Amoruso, A. Taddei, D. Holcman, Spatial telomere organization and clustering in yeast *Saccharomyces cerevisiae* nucleus is generated by a random dynamics of aggregation–dissociation. Mol. Biol. Cell **24**, 1791–1800 (2013)
21. S. Jacquot, A historical law of large numbers for the Marcus-Lushnikov process. Electron. J. Probab. **15**, 605–635 (2009)
22. F.P. Kelly, *Reversibility and Stochastic Networks*. Wiley Series in Probability and Mathematical Statistics (Wiley, Chichester, 1979)
23. P. Krapivsky, S. Redner, E. Ben-Naim, *A Kinetic View of Statistical Physics* (Cambridge University Press, Cambridge, 2010)
24. A.A. Lushnikov, Coagulation in finite systems. J. Colloid Interface Sci. **65**, 276–285 (1978)
25. A. Marcus, Stochastic coalescence. Technometrics **10**, 133–143 (1968)
26. H.G. Rotstein, Cluster-size dynamics: a phenomenological model for the interaction between coagulation and fragmentation processes. J. Chem. Phys. **142**, 224101 (2015)
27. H. Schober, et al., Controlled exchange of chromosomal arms reveals principles driving telomere interactions in yeast. Genome Res. **18**, 261–271 (2008)
28. Z. Schuss, *Theory and Applications of Stochastic Processes: An Analytical Approach*. Applied Mathematical Sciences, vol.170 (Springer, New York, 2010)
29. Z. Schuss, *Diffusion and Stochastic Processes: An Analytical Approach* (Springer, New York, 2010)
30. Z. Schuss, *Nonlinear Filtering and Optimal Phase Tracking*. Applied Mathematical Sciences, vol. 180 (Springer, New York, 2011)
31. P. Therizols, T. Duong, B. Dujon, C. Zimmer, E. Fabre, Chromosome arm length and nuclear constraints determine the dynamic relationship of yeast subtelomeres. Proc. Natl. Acad. Sci. U. S. A. **107**, 2025–2030 (2010)
32. C.J. Thompson, *Classical Equilibrium Statistical Mechanics* (Oxford University Press, Oxford, 1988)
33. B.R. Thomson, Exact solution for a steady-state aggregation model in one dimension. J. Phys. A **22**, 879–886 (1989)
34. M. von Smoluchowski, Drei Vorträge über Diffusion Brownsche Molekularbewegung und Koagulation von Kolloidteichen. Phys. Z. **17**, 557–571 (1916)
35. J.A. Wattis, An introduction to mathematical models of coagulation–fragmentation processes: a discrete deterministic mean-field approach. Physica D **222**, 1–20 (2006)
36. R. Yvinec, M.R. D'Orsogna, T. Chou, First passage times in homogeneous nucleation and self-assembly. J. Chem. Phys. **137**, 244107 (2012)
37. A. Zlotnick, Theoretical aspects of virus capsid assembly. J. Mol. Recognit. **18**, 479–490 (2005)

A Review of Stochastic and Delay Simulation Approaches in Both Time and Space in Computational Cell Biology

Kevin Burrage, Pamela Burrage, Andre Leier, and Tatiana Marquez-Lago

1 Introduction

Heterogeneity is a key property of biological systems at all scales: from the molecular level to the population level [1–3]. Many systems have evolved ways of either minimising or exploiting this noise [4, 5]. In particular, some critical cellular systems have evolved to minimise or take advantage of noise [6, 7], for example, persistence in bacteria [8, 9]. We can classify heterogeneity as arising from three main sources: *genetic* (nature), *environmental* (nurture) and *stochastic* (chance). In this chapter we focus on the latter. We make the distinction between Environmental heterogeneity often called *extrinsic* noise and *intrinsic* noise, which arises from random thermal fluctuations [1, 9].

Intrinsic noise was perhaps first observed in a cell biology setting by Spudich and Koshland [10]. They noticed that individual bacteria from an isogenic population maintained different swimming patterns throughout their entire lives. Although they called this 'non-genetic individuality', they believed it arose from random fluctuations of low copy-number molecule or *intrinsic* noise. This affects the microscopic DNA, RNA and protein molecules in many ways, most notably via the Brownian motion of molecules and the randomness associated with their reactions. It ensures that the process of gene expression, one of the most fundamental processes of a cell, proceeds differently each time, *even in the absence* of the previous two sources of heterogeneity [9, 11].

K. Burrage (✉) • P. Burrage
ARC Centre of Excellence for Mathematical and Statistical Frontiers, School of Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, QLD, 4000, Australia
e-mail: kevin.burrage@qut.edu.au

A. Leier • T. Marquez-Lago
Department of Genetics and Informatics Institute, University of Alabama at Birmingham, School of Medicine, Birmingham, AL, 35294, USA

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_11

Some biological systems have evolved to make use of intrinsic noise: a good example is persister-type bacteria, which can withstand antibiotic treatments even though they do not have genetic mutations for resistance [12]. These effects can arise through the phenomenon of stochastic switching, where cells randomly transition from one state to another [13, 14]. Another well-known example of exploiting stochasticity is the bacteriophage Lambda decision circuit [15]. Stochasticity also plays a crucial role in causing genetic mutations [16, 17]; these are essential for creating the heritable heterogeneity upon which natural selection can act, thus allowing evolution to occur [18].

On the other hand, intrinsic noise can interfere with the precise regulation of molecular numbers [19], and cells have to compensate for this by adopting mechanisms that enable robustness of certain key properties [4, 5], such as feedback loops [9, 11]. Furthermore, mutations may also have negative effects, as certain mutations in, for example, stem and somatic cells are thought to be the cause of cancer [20–22].

2 Temporal Modelling in Computational Cell Biology

There has been a long and successful history in computational cell biology of using rate kinetic ordinary differential equations to model chemical kinetics within a living cell. For instance, these techniques have been applied on the plasma membrane, in the cytosol and in the nucleus of eukaryotic cells to understand cell processes ranging from gene regulation to transport between cellular compartments. Modifications via delay differential equations were first considered as early as [23], in order to represent the fact that the complex regulatory processes of transcription and translation were not immediate but were in fact examples of delayed processes.

Thus in a purely temporal homogeneous setting and when there are large numbers of molecules present, chemical reactions are modelled by ordinary differential equations that are based on the Law of Mass Action and that estimate reaction rates on the basis of average values of the reactant density. Any set of m chemical reactions can be characterised by two sets of quantities: the stoichiometric vectors (update rules for each reaction) $\nu_1, \ldots, \nu_m$ and the propensity functions $a_1(X(t)), \ldots, a_m(X(t))$. The propensity functions represent the relative probabilities of each of the m reactions occurring. These are formed by multiplying the rate constant and the product of the reactants on the left-hand side of each reaction. Here $X(t)$ is the vector of concentrations at time t of the N species involved in the reactions. The ODE that describes this chemical system is

$$X'(t) = \sum_{j=1}^{m} \nu_j a_j\left(X(t)\right).$$

This formulation may have many different timescales (stiffness) but there are a wide variety of numerical methods that can deal effectively with such systems.

It was the pioneering work of Gillespie [24] and Kurtz [25] who challenged this deterministic view of cellular kinetics. They argued that when the cellular environment contained small to moderate numbers of proteins, then the Law of Mass Action is not an adequate description of the underlying chemical kinetics as it only describes the average behaviour. In this regard, the fundamental underlying principle is that of intrinsic noise. Intrinsic noise is associated with the inherent uncertainty in knowing when a reaction occurs and what that reaction is. The variance associated with this uncertainty increases as the number of proteins in the cellular environment becomes small. [24, 25] showed how to model intrinsic noise through the concept of nonlinear discrete Markov processes and Poisson processes, respectively. These two approaches both model the same processes and are now lumped together under the title the Stochastic Simulation Algorithm (SSA), although their formulations are different. The essential observation underlying the SSA is that the waiting time between reactions is exponentially distributed and that the most likely reaction to occur in this time interval is based on the relative sizes of the propensity functions.

More formally, the SSA is an exact procedure that describes the evolution of a discrete nonlinear Markov process. It accounts for the inherent stochasticity of the m reactions within a system and only assigns integer numbers of molecules to the state vector. At each step, the SSA samples two random numbers from the uniform distribution U[0,1] to evaluate an exponential waiting time, τ, for the next reaction to occur and an integer j between 1 and m that indicates which reaction occurs. The state vector is updated at the new time point by the addition of the j^{th} stoichiometric vector to the previous value of the state vector, that is,

$$X(t+\tau) = X(t) + \nu_j.$$

Several more efficient, but more complex, variants of the SSA have been developed [26–27]. However, despite these increases in computational speed, the SSA has an inherent limitation: it must simulate every single reaction, and in cases where there are many reactions or molecular populations become too large, it is computationally intensive.

In a slightly different vein, the SSA describes the evolution of a nonlinear discrete Markov process and as such this stochastic process has a probability density function whose solution is described by the Chemical Master Equation (CME). The CME is a discrete parabolic partial differential equation in which there is one equation for each configuration of the 'state space'. When the state space is enumerated, the CME becomes a linear ODE and the probability density function takes the form

$$p(t) = e^{At}p(0).$$

Here A is the state-space matrix. Thus the solution of the CME can be reduced to the computation of the evolution of the exponential of a matrix times an initial

probability vector. As there is one equation for each possible configuration of the state space this can be very computationally challenging, although recently developed methods can cope with some of these computational costs [28–32] to make this a very feasible technique.

The main limiting feature of SSA is that the time step can become very small, especially if there are large numbers of molecules or widely varying rate constants. Thus tau-leap methods have been proposed in which the sampling of likely reactions is taken from either Poisson [33] or Binomial [34] distributions. In these approaches a much larger time step can be used at the loss of a relatively small amount of accuracy. The tau-leap method [33] allows steps that are much larger than the SSA by estimating the *total number of occurrences* of each type of reaction over that step. Thus if there are m reactions, we take m Poisson random number samples based on the sizes of the propensity functions evaluated at the beginning of the step. The algorithm can thus be written as

$$X(t+\tau) = X(t) + \sum_{j=1}^{m} \upsilon_j P(\tau a_j(X(t))) .$$

Note that now we have complete flexibility in choosing τ and it can be done adaptively in such a way as to control the local error at each step. This approach has the key advantage that individual reactions need not be simulated. However, the main drawback is a loss of accuracy compared to the SSA as a function of the step size. Several schemes have been devised to optimise the time step [35, 36], while some implementations combine the SSA with tau-leaping via a threshold on the number of molecules in the system at any given time. However, it is possible that molecular species that are depleted in any reaction can go negative if the time step is too large and schemes have been developed that allow the choice of larger time steps whilst avoiding negative populations [34, 36–40]. One important way in which this can be done is to sample from the Binomial distribution rather than the Poisson distribution [34]. Another approach is to consider the order of accuracy as a function of the step size. In this regard the (weak) order of accuracy of the tau-leap method can be shown to be one [41–43], meaning that error decreases proportionally to the time step. Higher-order methods allow larger time steps to achieve the same error, thus decreasing computational time [43–47].

Although it is not uncommon for chemical systems to be rather complicated, a difficult situation arises when a system has reactions that operate at very different timescales, for instance, slow and fast. Although standard tau-leap and higher-order tau-leap methods are able to simulate these systems, their time steps must be reduced, which can dramatically slow down simulation time when the separation of the scales is significant and the fast reactions occur frequently.

In such cases, there are two options: either to use special methods for these ‘stiff’ systems or to use multiscale methods. The former are often based on deterministic methods for stiff ordinary differential equation systems and allow the use of normal time steps by expanding the range of time steps where the method is stable, thus

opening the door for stiff systems to be simulated in similar time periods as non-stiff ones [48–50]. Multiscale methods, on the other hand, partition the reactions into fast and slow types, and simulate the fast reactions using an approximate method and the slow reactions using an accurate stochastic method. The partitioning implies that the slow reactions are constant over the timescale of the fast reactions and that the fast reactions have relaxed to asymptotic values between each slow reaction. Thus, the two sets of reactions are simulated iteratively to take into account the coupling. This can also be generalised to three regimes: fast, medium and slow [51]. This approach allows for very significant reductions in computational time as the fast reactions, which would take up the most computational effort, can be simulated very quickly with continuous methods. However, it also introduces errors associated with coupling between two scales, as well as the possibility of errors from the simulation of the fast reactions. An interesting approach runs short bursts of a single SSA for the *fast* reactions, which is used to infer parameters for a differential equation approximation of the slow reactions [52, 53].

There is in fact an intermediate regime that can still capture the inherent stochastic effects but reduce the computational complexity associated with the SSA. This intermediate framework is called the Chemical Langevin Equation (CLE). It is described by an Itô stochastic differential equation (SDE) driven by a set of Wiener processes that describes the fluctuations in concentrations of the molecular species. Various numerical methods can then be applied to this equation—the simplest method being the Euler–Maruyama method [54].

The CLE attempts to preserve the correct dynamics for the first two moments of the SSA and takes the form

$$dX = \sum_{j=1}^{m} \nu_j a_j\left(X(t)\right) dt + B\left(X(t)\right) dW(t).$$

Here $W(t) = (W_1(t), \ldots, W_N(t))^{\mathrm{T}}$ is a vector of N independent Wiener processes whose increments $\Delta W_j = W_j(t+h) - W_j(t)$ are $N(0,h)$ and where

$$B(x) = \sqrt{C}, \quad C = (\nu_1, \ldots, \nu_m)\,\mathrm{Diag}\left(a_1(X), \ldots, a_m(X)\right)(\nu_1, \ldots, \nu_m)^{\mathrm{T}}.$$

Here h is the time discretisation step. Effective methods designed for the numerical solution of SDEs [54–56] can be used to simulate the chemical kinetics in this intermediate regime. [57] have shown how to construct the CLE so that it minimises the number of Wiener processes. Furthermore, adaptive multiscale methods have been developed that attempt to move back and forth between the deterministic and stochastic regimes as the numbers of molecules change [51].

These temporal approaches are applied under the hypothesis of homogeneous and well-mixed systems. It is well known, for example that diffusion on the cell membrane is not only highly anomalous but the diffusion rate of proteins on live cell membranes is also between one and two orders of magnitude slower than in reconstituted artificial membranes with the same composition [58]. Furthermore,

diffusion is dependent on the dimensions of the medium so that diffusion on the highly disordered cell membrane is not a perfectly mixed process and therefore the assumptions underlying the classical theory of chemical kinetics fail, requiring either new approaches to modelling chemistry on a spatially crowded membrane [59] or methods based on detailed spatial simulations.

However, rather than abandoning temporal models entirely, it is possible to capture important spatial aspects and incorporate them into temporal models. This can be done in a number of ways. For example, compartmental models have been developed that couple together the plasma membrane, cytosol and nucleus—see, for example [60], in which an SSA implementation of Ras nanoclusters on the plasma membrane is coupled with an ODE model for the MAPK pathway in the cytosol. Diffusion and translocation can be captured through the use of distributed delays that can then be incorporated into mathematical frameworks through the use of delay differential equations or delay variants of the Stochastic Simulation Algorithm (see [61], for example). [62] have explored a number of spatial scenarios, run detailed spatial simulations to capture diffusion and translocation processes and then incorporated this information into purely temporal models through distributed delays—see Sect. 5 for more details. Another way in which spatial information can be captured and then incorporated into purely temporal models is the area of anomalous diffusion, where spatial crowding and molecular binding can affect chemical kinetics. In this setting the mean square deviation of a diffusing molecule is no longer linear but sublinear in time t and of the form

$$E\left[X^2(t)\right] = 2Dt^{\alpha}, \quad \alpha \in (0, 1].$$

Here, α is called the anomalous diffusion parameter. If the value of α can be estimated, either experimentally or from detailed Monte Carlo simulations, then the SSA can be modified so that the waiting time between reactions is no longer exponentially distributed but has a heavy tail [59].

3 Spatial Models in Computational Cell Biology

One of the fundamental goals of integrative Computational Biology is to understand complex spatio-temporal processes within cells. However, such a task may become exceedingly difficult, due to the intrinsic multiscale nature of these processes. For example, in order to fully understand cell signal transduction, a careful description of membrane-bound receptor activation processes must be accounted for. However, the plasma membrane is a highly complex structure, compartmentalised on multiple length and timescales stemming from lipid–lipid, lipid–protein interactions with the cytoskeleton. As a result, detailed simulations accounting for all processes may be computationally prohibitively expensive, or simply infeasible.

Of the class of spatial methods, the approach with the lowest computational cost consists of solving reaction-diffusion partial differential equations, representing the concentration of a given molecular species in the system. However, this approach is only valid if and when: (1) all molecular species in the system have large molecular concentrations, and (2) noise is not amplified throughout the system. If at least one of these conditions fails to hold, we must rely on spatial stochastic simulation that can be discrete or continuous in form. In the discrete spatial setting either lattice or off-lattice particle based methods are appropriate.

As a particular example of lattice based methods, we consider the plasma membrane. The plasma membrane is a complex and crowded environment that has many roles including signalling, cell–cell communication, cell feeding and excretion and protection of the interior of a cell. It is heterogeneous—the cytoskeletal structure just inside the plasma membrane can corral and compartmentalise membrane proteins. Chemically inert objects can form barriers to protein diffusion on the plasma membrane. Trying to capture such complexity using higher-level mathematical frameworks such as partial differential equations is extremely challenging, so instead a stochastic spatial model using Monte-Carlo simulation is appropriate. In such an approach the plasma membrane can be mapped to a two dimensional lattice, usually regular but not necessarily so. The size of each computational cell 'voxel' depends on the biological questions that are being addressed, but taking into account volume-exclusion effects, usually the voxel is such that at most one protein per voxel is allowed. A protein undergoes a random walk, so that at each time step a protein is selected at random, and a movement direction (north, south, east or west, in the case of a rectangular lattice arrangement) is randomly determined. The distance moved depends on the diffusion rates for each species. Chemical reactions can be simulated by checking the chemical reaction rules and then replacing that protein and/or creating a new protein at that location whenever a collision (volume exclusion event) occurs [63, 64]. Although only relatively small sections of the membrane on short timescales can be simulated with this approach due to the very slow computational performance, we note that diffusion can be considered as a unimolecular reaction. Thus if we order the voxel elements within the lattice into a vector, we can consider this method in the SSA framework and apply the same tools that have been developed in the purely temporal setting. This can allow us to make use of vectorisation to speed up the performance. Furthermore, there is a spatial CME associated with this approach [65] and again techniques used in the purely temporal case can be used, although at the cost of very significantly increased computational complexity.

In off-lattice methods, all particles in the system have explicit spatial coordinates at all times. At each time step, molecules are able to move, in a random walk fashion, to new positions. In many cases, reaction zones whose size depends on the particular diffusion rates are drawn around each particle. If one or more molecules happen to be inside such a zone, appropriate chemical reactions can take place with a certain probability; if a reaction is readily performed, the reactant particles are flagged, to avoid repetition of chemical events. Noticeably, in off-lattice methods, the domains and/or compartments can still be discretised, to aid the localisation of particles

within the simulation domain. In this vein, [66] have considered how to combine tau-leaping and compartmentalisation in a spatial setting.

A less computationally intensive alternative, albeit still costly in many scenarios, is to consider molecular interactions in the mesoscopic realm. Here, the discretisation of the Reaction-Diffusion Master Equation (RDME) results in reactive neighbouring subvolumes within which several particles can coexist, with well-mixedness assumed in each subvolume. There are a number of algorithms extending discrete stochastic simulators to approximate solutions of the RDME by introducing diffusion steps as first-order reactions, with a reaction rate constant proportional to the diffusion coefficient. In [67, 68] the authors provide the specific outline for extending discrete stochastic simulators to the RDME regime, while the algorithms in [69, 70] provide clever extensions of the 'next reaction method' [26], commonly known as the 'next subvolume method'. A review on the construction of such methods is given in [71].

The Next Subvolume Method [69, 70, 72, 73] is a generalisation of the SSA [24], where the simulation domain is divided into uniform separate subvolumes that are small enough to be considered homogeneous by diffusion over the timescale of the reaction. At each step, the state of the system is updated by performing an appropriate reaction within a single subvolume or by allowing a molecule to jump to a randomly selected neighbouring subvolume. Diffusion is then modelled as a unary reaction. In a two dimensional setting the rate is proportional to the diffusion coefficient divided by the length of a side of the subvolume. In this way, diffusion inside the algorithm becomes another possible event with a regular propensity function, and follows the same update procedure as any chemical reaction. The expected time for the next event in a subvolume is calculated in a similar way to the SSA algorithm, including the reaction and diffusion propensities of all molecules contained in that subvolume, at that particular time. However, times for subsequent events will only be recalculated for those subvolumes that were involved in the current time step, and they are subsequently re-ordered in an event queue. Similar to accelerated approaches to simulate exact trajectories from the CME, there exist methods to coarse-grain, and therefore accelerate, computations for the RDME [66]. Separately, [74] split the time integration of the RDME into a macroscopic diffusion (for species with large numbers of molecules) and a stochastic mesoscopic reaction/diffusion part (for species with small numbers) obtaining the mesoscopic diffusion coefficients from appropriate Finite Element discretisations.

In addressing these spatial issues, we are led to consider the role of anomalous diffusion. Anomalous diffusion refers to processes where the mean squared displacement (MSD) of a particle is no longer linear in time [75]. Anomalous diffusion can be viewed in two ways: as a mechanism to localise molecules and control signalling [76], or a macroscopic result of underlying microscopic events. From the experimental perspective, various techniques have been used to study such processes, including Single Particle Tracking [77], Fluorescence Recovery after Photo bleaching [78] and Fluorescence Correlation Spectroscopy [79]. However, the quantification of the degree and nature of the anomalous diffusion has shown to be more difficult than anticipated, due to experimental limitations [76, 80]. In spite

of these discussions and recent technical developments, the nature of anomalous subdiffusion is still not well understood, either from experimental or simulation perspectives.

There are many reasons for these discrepancies including the fact that often only short tracks are recorded and the MSD relationship is not necessarily a robust metric for inferring complex spatial information. A more robust metric would be to construct a probability density function evolving in space and time, but this would require a very time-consuming experimental study. From the simulation perspective, a number of simulation studies have reported crowding-dependent values of α [64, 81]. But a prediction of percolation theory is that for immobile obstacles α depends only on the embedding space dimension and not on the obstacle density [82]. From theoretical perspectives, in the case of immobile objects, we can expect to observe Brownian diffusion for initial small time periods (when no obstacles have yet been encountered), an anomalous regime at intermediate times and Brownian diffusion in the asymptotic regime [83, 84]. It is these crossovers that can create confusion when trying to interpret diffusive characteristics.

It remains to be seen whether anomalous diffusion scenarios can only be captured by explicit spatial models, or whether equivalent dynamics could be obtained using off-lattice simulations with single molecules devoid of explicit obstacles, thus capturing essential dynamics while significantly reducing computational time. Other possibilities would include deriving an SSA and its associated CME that would work in an anomalous diffusion temporal setting only. We could then attempt to replicate the deterministic and stochastic regimes in the time anomalous setting. This would lead to time fractional or space fractional representations. For example, molecular concentration dynamics $C(x, t)$ in a typical subdiffusive setting could be represented by the time-fractional differential equation

$$\frac{\partial}{\partial t} C\left(x, t\right) = K_\alpha D_t^{1-\alpha}\, \nabla^2 \, \mathrm{C}\left(x, t\right) + f\left(x, t\right),$$

where x and t are the space and time variables, respectively, and the anomalous exponent α is the fractional order of the time derivative operator. Here $D_t^{1-\alpha}\, g(t)$ is the fractional Riemann–Liouville derivative operator that reduces to the identity operator when $\alpha = 1$ while K_α is the fractional equivalent to the classical diffusion coefficient and has dimension $[K_\alpha] = [l]^2[t]^{-\alpha}$.

4 Modelling and Simulating Stochastic ion Channel Dynamics

Ion channels are multiconformational proteins that form a pore in the membrane of excitable cells. They open and close due to conformational changes in the proteins as a result of variations in membrane potential, and thus regulate the movement of ions across the lipid bilayer [85]. The dynamics of these proteins

are fundamental to the generation of an action potential (AP) in excitable cells [86]. Single-channel recordings have demonstrated that the conformational changes the protein undergoes as it opens and closes occur at random [87]. This internal stochasticity causes fluctuations in individual ionic conductances, [86], and has important effects on the electrical dynamics of the cell [88–91].

In neuronal cells, stochastic ion channel behaviour can modify a number of electrical properties of the cell including the firing threshold [92] and spike timings [93]. In cardiac myocytes this intrinsic randomness leads to variability in the duration of successive APs [89, 90], termed beat-to-beat variability, which is thought to be an indicator of arrhythmias [94]. It can even cause alterations to the AP morphology under pathological conditions, resulting early-after-depolarisations (EADs) [89].

While a discrete-state modelling and simulation approach is often seen as the 'gold-standard', it becomes increasingly computationally costly as the number of channels increases beyond a few hundred [95]. This has led to the increasing popularity of using stochastic differential equations to describe ion channel dynamics [89, 96, 97]. Fox and Lu [98] were the first to take such an approach, which they applied to the Hodgkin–Huxley model, and their method has since been extensively used to model neuronal cells [96, 97], cardiac myocytes [89, 90] and pancreatic beta cells [91]. Their approach reduces the dynamics of the whole channel to the collective dynamics of a series of gating variables that can be either open or closed. The proportion of each gating variable in the open state is described by a Stochastic Differential Equation (SDE) and the proportion of open channels is given by the product of the number of open gates.

However, a number of studies have demonstrated discrepancies between the SDE and discrete-state Markov chain model [95, 99, 100] [101]. Goldwyn et al. [101] demonstrate that constructing the SDE model in terms of the kinetic dynamics of the channel, rather than the individual gating variables, preserves the stochastic behaviour of the discrete-state Markov chain model, see the recent review [104] for a further discussion.

In order to illustrate these ideas we give more details on the form of the Chemical Langevin equation for ion channel dynamics. We first consider an ion channel transitioning just between the closed and open states. Let the proportion of channels in the open state be y, let N be the total number of ion channels and W be a Wiener process. Then the form of the CLE is

$$dy = (a - (a + b)\, y)\, dt + \frac{1}{\sqrt{N}} \sqrt{a + (b - a)\, y} dW.$$

However, the complicating factor is that the parameters a and b are themselves functions of the trans-membrane voltage. In the more general setting when there are a number of transitions between the various ion channel states then the formulation is given by

$$dy(t) = vD\,(y(t))\, dt + \frac{1}{\sqrt{N}} \sum_{p=1}^{d/2} b^p\,(y(t))\, dW_p + k_t.$$

For the moment take $k = 0$. Here the columns of v are state changes resulting from a transition. D is a diagonal matrix whose entries are chance of transition occurring (namely the propensity functions), $b^p(y(t))$ are columns of matrix B, where $\mathrm{BB}^{\mathrm{T}} = vD(\mathbf{y}(t))v^{\mathrm{T}}$. In the case of, say p, unimolecular transitions, which is the setting for ion channels, then this simplifies to $\mathrm{B} = ES$, where E is a $p \times \mathrm{d}/2$ matrix and S is a $\mathrm{d}/2 \times \mathrm{d}/2$ diagonal matrix. In particular, E is a matrix with 1s on the diagonal, -1s in certain lower triangular positions and 0s elsewhere. Here S is a diagonal matrix with square roots of linear combinations of certain pairs of components of y [57].

For example, in the case of Sodium and Potassium ion channels, that form key elements of the Hodgkin-Huxley ion channel model [103], the transitions are given by

$$
\begin{array}{ccccccccc}
n_0 & \underset{b_n}{\overset{4a_n}{\rightleftharpoons}} & n_1 & \underset{2b_n}{\overset{3a_n}{\rightleftharpoons}} & n_2 & \underset{3b_n}{\overset{2a_n}{\rightleftharpoons}} & n_3 & \underset{4b_n}{\overset{a_n}{\rightleftharpoons}} & n_4 \\
m_0h_0 & \underset{b_m}{\overset{3a_m}{\rightleftharpoons}} & m_1h_0 & \underset{2b_m}{\overset{2a_m}{\rightleftharpoons}} & m_2h_0 & \underset{3b_m}{\overset{a_m}{\rightleftharpoons}} & m_3h_0 & & \\
b_h \uparrow\downarrow a_h & & b_h \uparrow\downarrow a_h & & b_h \uparrow\downarrow a_h & & b_h \uparrow\downarrow a_h & & \\
m_0h_1 & \underset{b_m}{\overset{3a_m}{\rightleftharpoons}} & m_1h_1 & \underset{2b_m}{\overset{2a_m}{\rightleftharpoons}} & m_2h_1 & \underset{3b_m}{\overset{a_m}{\rightleftharpoons}} & m_3h_1. & &
\end{array}
$$

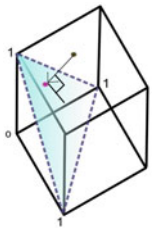
In the case of Sodium, for example, then E and S are given by

$$
E = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad \mathrm{diag}(S) = \begin{pmatrix} \sqrt{a_n y_{n_3} + 4b_n y_{n_4}} \\ \sqrt{2a_n y_{n_2} + 3b_n y_{n_3}} \\ \sqrt{3a_n y_{n_1} + 2b_n y_{n_2}} \\ \sqrt{4a_n y_{n_0} + b_n y_{n_1}} \end{pmatrix}.
$$

In passing, we note that due to the structure of the ion channels there is an explicit formulation to the underlying probability density function in terms of multinomials [104] but as the transition rates are nonlinear functions of the voltage this does not really help in simulating the ion channel dynamics.

Although this approach gives the correct dynamics, the solution to the SDE must remain non-negative for the path to have any biological relevance. Yet it has been shown that the solution can become negative [105]. Furthermore, since the noise term in the SDE model involves the square root of some function of the state variable, this can result in numerical solutions becoming imaginary [105]. Alterations to the numerical scheme can be made to force the solution to remain positive, for example, the Wiener increment can be continually resampled [89]. However, such alterations can bias the results [105–107]. Another approach is to replace the variable in the noise term with its equilibrium value, [96, 101] so that the square root term is independent of the state of the system. However, such an approach can still result in the proportion of channels becoming negative [101]. In [105] a hybrid simulation method for the Hodgkin–Huxley model is developed that

attempts to improve the computational efficiency of the discrete-state Markov chain model whilst ensuring individual simulations remain non-negative by switching between regimes.



Example of the reflection process on a three dimensional bounded domain

We argue that the boundary conditions are not naturally incorporated into the standard CLE and that they can do so by the use of reflected SDEs [108] in which k_t in the general formulation is a reflected process that only comes into play at the boundaries of the domain. The basic idea is to evaluate the non-reflected process at the next time step using the Euler–Maruyama method. If the simulation is in the desired region, set Y at the next time step to be this value. Otherwise orthogonally project onto the boundary of the domain D. We assume that the process will reflect $y(t)$ into the interior of the domain in the direction of the inward pointing unit normal. It can be shown that this approach will converge with strong order ½-h for all $h > 0$. This can be visualised for the three state models—see figure. Thus in our general formulation we can interpret k_t as a reflecting process and this determines the behaviour at the boundary.

Finally, very recently Schnoerr et al. [109] show in a very nice paper that, by extending the domain of the CLE to the Complex space, the CLE's accuracy for unimolecular systems is restored. This is at the cost of having to perform simulations in complex arithmetic and taking care with the use of pre-defined functions.

5 The Role of Delays in Modelling Biochemical Reaction Systems and Model Reduction

The origins of delay differential equations (DDEs) date, most likely, back to the second half of the eighteenth century. According to Schmidt [110], some of the early work on DDEs (and, more generally, the so-called functional differential equations) originates from famous mathematicians such as Laplace, Poisson and Lacroix (see references in [110]). Two of the earliest examples of DDEs from the early twentieth century found in the mathematical medicine and mathematical biology literature are by Lotka, who studied a model of malaria epidemics with incubation delays (in particular in [111]), and by Volterra, who investigated delayed predator–prey models [112]. Ever since, DDEs have become an integral part of mathematical modelling of

biological, biomedical and physiological processes. Examples of delay models can be found in areas such as population dynamics and epidemiology [113, 114], gene regulation [115], cell signalling [116], viral dynamics [117], tumour growth [118], drug therapy [118, 119], immune response [120] and respiratory systems [121].

The use of delays and systems' histories was driven by the desire for more realistic and, consequently, more accurate mathematical models. Indeed, introducing delays is many times essential for reconciling models with observations and experimental data. Moreover, delays provide a way for a more intuitive modelling approach, i.e. phenomenological rather than mechanistic. In this case, complex processes are lumped while underlying mechanisms and inherent intermediate steps are not explicitly accounted for. Yet, the time that such processes require is included in the form of a constant delay or delay distribution.

With growing interest in the stochastic dynamics of chemical reaction networks, delays have also been introduced into stochastic simulation algorithms (SSAs). Several so-called delay SSAs (DSSAs) have been proposed [61, 122–125], and the Delay Chemical Master Equation (DCME) was introduced as a generalisation to the Chemical Master Equation for reactions with associated constant delays [61, 126] or delay distributions [126]. However, contrary to CMEs, for which either closed solutions have been presented or a large number of numerical and computational strategies have been suggested [104, 127–132], DCMEs have been largely disregarded. This is because DCMEs do no longer represent Markovian processes, since transitions between states depend on the current state and the process history. This leads to terms in a DCME that, for a delay reaction with stoichiometric vector ν and constant delay τ, look like

$$\sum_{x_i \in l(x)} \left(a\left(x_i\right) P\left(x - \nu, t; x_i, t - \tau\right) - a\left(x_i\right) P\left(x, t; x_i, t - \tau\right) \right),$$

where the sum is taken over all previous states x_i prior to the current state x. The joint probabilities are usually unknown and these terms can only be simplified if (a) the coupling of the system states at times t and $t - \tau$ is weak, in which case we obtain a reasonably good approximation; or (b) the triggering of the delayed reaction is fully independent of the occurrences of other reactions and of the state x_i at the time of triggering. The latter implies in particular that none of the reactions in the system, including the delayed reaction, change the number of reactants of the delayed reaction, nor its kinetic function [126]. In this case, the propensity function of the delayed reaction is constant, i.e. $a(x_i) = c$ for all x_i, and the above sum simplifies to

$$\begin{aligned} & c \sum_{x_i \in l(x)} \left(P\left(x - \nu, t | x_i, t - \tau\right) P\Big(x_i, t - \tau\Big) - P\big(x, t | x_i, t - \tau\big) P\big(x_i, t - \tau\big) \right) \\ & \quad = c \left(P\left(x - \nu, t\right) - P\Big(x, t\Big) \right). \end{aligned}$$

If the delay is given in the form of a distribution instead of a constant delay, then its cumulative density function appears as just another factor. The DCME remains a

homogeneous system of linear first-order ODEs, except that it now includes a time-dependent coefficient, and can then be solved numerically using the available tools for CMEs.

Moreover, the analytic solution for CMEs of monomolecular reaction systems was shown to be the convolution of multinomial and product Poisson distributions [104]. For a simple delayed, unidirectional reaction scheme, a multinomial distribution was recently derived as its solution using purely probabilistic arguments [126]. This suggests that such a distribution may also be the solution to a more general class of monomolecular, delayed reaction systems.

As already mentioned above, delays can also be used for model reduction. A number of reduction techniques have been proposed in the past, including the classical equilibrium approximation by Michaelis and Menten, the quasi-steady state approximation (QSSA) by Briggs and Haldane, several variations of the QSSA [133, 134], methods based on the linear noise approximation [135, 136], and the finite-state projection method [130]. However, they all approximate the true solution and/or make certain assumptions, for instance, a timescale separation, which, if not met, can lead to inaccurate results. The method presented in [137, 138] uses delayed reactions between species of interest that replace a number of intermediate species and reactions between these. The individual delays are obtained as first-passage time distributions in analytic form. These can then be placed into DSSA implementations for generating sample trajectories of the abridged system's dynamics. This method is fully accurate for all bidirectional, unimolecular reaction chains, including degradations, synthesis and bypass reactions, and allows for large computational savings. This holds true in particular if, over the same simulation time span, the number of reactions in the unabridged system is considerably larger than the number of DSSA steps in the abridged system.

Lastly, it has become evident that spatial aspects play a crucial role in biological and biomedical processes. Even in relatively simple biochemical systems, the observed behaviour can vary considerably from the often assumed, well-mixed scenario, where spatial dependencies, geometries and structures are not taken into account. Thus, it is important to consider these spatial aspects in modelling approaches, both for our understanding and for accurate predictions of such processes. While detailed spatial models are more realistic they are also much more computationally demanding—if not prohibitively expensive. Alternatively, we can try to incorporate the effects of such spatial features into temporal models, without any explicit spatial representation. As has been shown recently, delay distributions may provide an appropriate tool [62]. The proposed methodology is similar to the model reduction with delays described above: the first step consists of obtaining proper delay distributions. These can stem from diffusion profiles and can be directly obtained from particle simulations, in vitro experiments or corresponding PDE solutions. Such tailored distributions are then used along with their associated reactions in a modified version of the DSSA. When applied to a variety of simple scenarios of molecular translocation processes large computational savings were achieved.

6 Conclusions

There is still considerable ongoing research to refine both the SSA and approximate methods. However, there are four areas where significant improvements can be made easily. Firstly, graphical processing units on graphics cards are now very powerful for some applications. Parallelised stochastic methods allow us to harness this power in order to run many thousands of simulations simultaneously. Secondly, multiscale problems are common in many real applications and they are often large in nature. The use of adaptive multiscale approaches where processes on many different scales are integrated into one model could, for example, play an important role in personalised medicine. Thirdly, we can attempt to reduce the number of simulations needed in order to gain a certain accuracy using ideas from multi-level theory developed by Mike Giles—this is a form of variance reduction. Fourthly, the limitation of non-spatial methods is that they can only be accurately applied to macroscopically homogeneous systems, but this assumption does not hold in many (or even most) cases of interest. Therefore it is important to develop methods that can take appropriate account of such heterogeneous environments. Finally, we note that one area that we have not covered in this chapter is parameter inference and model selection. This area involves using statistical methods to find model parameters from experimental data, and to discriminate between those models best fit this data [139–141]. This is typically not an easy task, as the data may be noisy, missing or sparse; Bayesian approaches offer a way of addressing these issues [142].

Acknowledgements The authors would like to thank the Isaac Newton Institute for Mathematical Sciences at the University of Cambridge, UK under the scientific program entitled Stochastic Dynamical Systems in Biology: Numerical Methods and Applications for supporting our visit and participation.

References

1. H.H. McAdams, A. Arkin, It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet. **15**, 65–69 (1999)
2. S. Huang, Non-genetic heterogeneity of cells in development: More than just noise. Development **136**, 3853–3862 (2009)
3. S.V. Avery, Microbial cell individuality and the underlying sources of heterogeneity. Nat. Rev. Microbiol. **4**, 577–587 (2006)
4. N. Barkai, B. Shilo, Variability and robustness in biomolecular systems. Mol. Cell **28**, 755–760 (2007)
5. C.V. Rao, D.M. Wolf, A.P. Arkin, Control, exploitation and tolerance of intracellular noise. Nature **420**, 231–237 (2002)
6. H.B. Fraser, A.E. Hirsh, G. Giaever, J. Kumm, M.B. Eisen, Noise minimization in eukaryotic gene expression. PLoS Biol. **2**, 834–838 (2004)
7. B. Lehner, Selection to minimise noise in living systems and its implications for the evolution of gene expression. Mol. Syst. Biol. **4**, 170 (2008)

8. D. Fraser, M. Kaern, A chance at survival: Gene expression noise and phenotypic diversification strategies. Mol. Microbiol. **71**, 1333–1340 (2009)
9. M. Kaern, T. Elston, W. Blake, J. Collins, Stochasticity in gene expression: From theories to phenotypes. Nat. Rev. Genet. **6**, 451–464 (2005)
10. J.L. Spudich, D.E. Koshland Jr., Non-genetic individuality: Chance in the single cell. Nature **262**, 467–471 (1976)
11. N. Maheshri, E.K. O'Shea, Living with noisy genes: How cells function reliably with inherent variability in gene expression. Annu Rev Bioph Biom **36**, 413–434 (2007)
12. N.Q. Balaban, J. Merrin, R. Chait, L. Kowalik, S. Leibler, Bacterial persistence as a phenotypic switch. Science **305**, 1622–1625 (2004)
13. M. Acar, J.T. Mettetal, A. van Oudenaarden, Stochastic switching as a survival strategy in fluctuating environments. Nat. Genet. **40**, 471–475 (2008)
14. P.J. Choi, L. Cai, K. Frieda, X.S. Xie, A stochastic single-molecule event triggers phenotype switching of a bacterial cell. Science **322**, 442–446 (2008)
15. A. Arkin, J. Ross, H.H. McAdams, Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. Genetics **149**, 1633–1648 (1998)
16. E.C. Friedberg, G.C. Walker, W. Siede, R.D. Wood, *DNA Repair and Mutagenesis*, 2nd edn. (ASM Press, Washington, DC, 2006)
17. J.M. Pennington, S.M. Rosenberg, Spontaneous DNA breakage in single living *Escherichia coli* cells. Nat. Genet. **39**, 797–802 (2007)
18. T.R. Gregory, Understanding natural selection: Essential concepts and common misconceptions. Evo Edu Outreach **2**, 156–175 (2009)
19. N. Rosenfeld, J.W. Young, U. Alon, P.S. Swain, M.B. Elowitz, Gene regulation at the single-cell level. Science **307**, 1962–1965 (2005)
20. L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, S.A.J.R. Aparicio, S. Behjati, et al., Signatures of mutational processes in human cancer. Nature **500**, 415–421 (2013)
21. M.R. Stratton, P.J. Campbell, P.A. Futreal, The cancer genome. Nature **458**, 719–724 (2009)
22. I.A. Rodriguez-Brenes, N.L. Komarova, D. Wodarz, Evolutionary dynamics of feedback escape and the development of stem-cell driven cancers. Proc. Natl. Acad. Sci. U. S. A. **108**, 18983–18988 (2011)
23. B.C. Goodwin, Oscillatory behavior in enzymatic control processes. Adv. Enzym. Regul. **3**, 425–438 (1965)
24. D.T. Gillespie, Exact stochastic simulation of coupled chemical reactions, J. Phys. Chemistry **81**(25), 2340–2361 (1977)
25. T.G. Kurtz, The relationship between stochastic and deterministic models for chemical reactions. J. Chem. Phys. **57**(7), 2976–2978 (1972)
26. M.A. Gibson, J. Bruck, Efficient exact stochastic simulation of chemical systems with many species and many channels. J. Phys. Chem. A **104**(9), 1876–1889 (2000)
27. Y. Cao, H. Li, L.R. Petzold, Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. J. Chem. Phys. **121**, 4059–4067 (2004)
28. S. MacNamara, A.M. Bersani, K. Burrage, R.B. Sidje, Stochastic chemical kinetics and the total quasi-steady-state assumption: Application to the stochastic simulation algorithm and chemical master equation. J. Chem. Phys. **129**(9), 095105 (2008)
29. S. MacNamara, K. Burrage, R.B. Sidje, Multiscale modeling of chemical kinetics via the master equation. SIAM J Multiscale Modelling and Simulation Multiscale **6**(4), 1146–1168 (2008)
30. S. Peleš, B. Munsky, M. Khammash, Reduction and solution of the chemical master equation using time scale separation and finite state projection. J. Chem. Phys. **125**, 204104-1–20410413 (2006)
31. T. Jahnke, S. Galan,*Solving chemical master equations by an adaptive wavelet method*, in *Numerical Analysis and Applied Mathematics: International Conference on Numerical Analysis and Applied Mathematics, 16–20 Sept*, ed. by T. E. Simos, G. Psihoyios, C. Tsitouras, vol. 1048 (AIP Conference Proceedings, Psalidi, Kos, Greece, 2008), pp. 290–293
32. S. Engblom, Galerkin spectral method applied to the chemical master equation. Commun Comput Phys **v5**(i5), 871–896 (2009)

33. D.T. Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems. J. Chem. Phys. **115**(4), 2001 (1716—1733)
34. T. Tian, K. Burrage, Binomial leap methods for simulation stochastic chemical kinetics. J. Chem. Phys. **121**, 10356–10364 (2004)
35. D.T. Gillespie, L.R. Petzold, Improved leap-size selection for accelerated stochastic simulation. J. Chem. Phys. **119**, 8229–8234 (2003)
36. Y. Cao, D.T. Gillespie, L.R. Petzold, Efficient step size selection for the tau-leaping simulation method. J. Chem. Phys. **124**, 044109 (2006)
37. A. Chatterjee, D.G. Vlachos, M.A. Katsoulakis, Binomial distribution based tau-leap accelerated stochastic simulation. J. Chem. Phys. **122**, 024112 (2005)
38. X. Peng, W. Zhou, Y. Wang, Efficient binomial leap method for simulating chemical kinetics. J. Chem. Phys. **126**, 224109 (2007)
39. M.F. Pettigrew, H. Resat, Multinomial tau-leaping method for stochastic kinetic simulations. J. Chem. Phys. **126**, 084101 (2007)
40. C.A. Yates, K. Burrage, Look before you leap: A confidence-based method for selecting species criticality while avoiding negative populations in tau-leaping. J. Chem. Phys. **134**, 084109 (2011)
41. M. Rathinam, L.R. Petzold, Y. Cao, D.T. Gillespie, Consistency and stability of tau-leaping schemes for chemical reaction systems. Multiscale Model Sim **4**, 867–895 (2005)
42. T. Li, Analysis of explicit tau-leaping schemes for simulating chemically reacting systems. Multiscale Model Sim **6**, 417–436 (2007)
43. Y. Hu, T. Li, B. Min, A weak second order tau-leaping method for chemical kinetic systems. J. Chem. Phys. **135**, 024113 (2011)
44. Y. Hu, T. Li, Highly accurate tau-leaping methods with random corrections. J. Chem. Phys. **130**, 124109 (2009)
45. D.F. Anderson, M. Koyama, Weak error analysis of numerical methods for stochastic models of population processes. Multiscale Model Sim **10**, 1493–1524 (2012)
46. T. Székely Jr., K. Burrage, R. Erban, K.C. Zygalakis, A higher-order numerical framework for stochastic simulation of chemical reaction systems. BMC Syst. Biol. **6**, 85 (2012)
47. Z. Xu, X. Cai, Unbiased tau-leap methods for stochastic simulation of chemically reacting systems. J. Chem. Phys. **128**, 154112 (2008)
48. M. Rathinam, L.R. Petzold, Y. Cao, D.T. Gillespie, Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. J. Chem. Phys. **119**, 12784–12794 (2003)
49. Y. Cao, D.T. Gillespie, L.R. Petzold, The adaptive explicit-implicit tau-leaping method with automatic tau selection. J. Chem. Phys. **126**, 224101 (2007)
50. P. Rué, J. Villa-Freixà, K. Burrage, Simulation methods with extended stability for stiff biochemical kinetics. BMC Syst. Biol. **4**, 110–123 (2010)
51. K. Burrage, T. Tian, P. Burrage, A multi-scaled approach for simulating chemical reaction systems. Prog Biophys Mol Bio **85**, 217–234 (2004)
52. R. Erban, I.G. Kevrekidis, D. Adalsteinsson, T.C. Elston, Gene regulatory networks: A coarse-grained, equation-free approach to multiscale computation. J. Chem. Phys. **124**, 084106 (2006)
53. S.L. Cotter, K.C. Zygalakis, I.G. Kevrekidis, R. Erban, A constrained approach to multiscale stochastic simulation of chemically reacting systems. J. Chem. Phys. **135**, 094102 (2011)
54. P.E. Kloeden, E. Platen, *Numerical Solution of Stochastic Differential Equations* (Springer-Verlag, Berlin, 1992)
55. P.M. Burrage, K. Burrage, A variable stepsize implementation for stochastic differential equations. SIAM J. Sci. Comput. **24**(3), 848–864 (2002)
56. P.M. Burrage, R. Herdiana, K. Burrage, Adaptive stepsize based on control theory for SDEs. J Comp and App Math **170**, 317–336 (2004)
57. B. Mélykúti, K. Burrage, K.C. Zygalakis, Fast stochastic simulation of biochemical reaction systems by alternative formulations of the chemical Langevin equation, J. Chem. Phys. **132**, 1 (2010)
58. K. Murase, T. Fujiwara, T.Y. Umemura, Ultrafine membrane compartments for molecular diffusion as revealed by single molecule techniques. Biophys. J. **86**, 4075–4093 (2004)

59. D.V. Nicolau, K. Burrage, Stochastic simulation of chemical reactions in spatially complex media. Computers & Mathematics with Applications **55**(5), 1007–1018 (2008)
60. T. Tian, A. Harding, E. Westbury, J. Hancock, Plasma membrane nano-switches generate robust high-fidelity Ras signal transduction. Nat. Cell Biol. **9**, 905–914 (2007)
61. M. Barrio, K. Burrage, A. Leier, T. Tian, Oscillatory regulation of Hes1: Discrete stochastic delay modelling and simulation. PLoS Comp Bio **2**(9), e117 (2006). doi:10.1371/journal.pcbi.0020117
62. T.T. Marquez-Lago, A. Leier, K. Burrage, Probability distributed time delays: Integrating spatial effects into temporal models. BMC Syst. Biol. **4**, 19 (2010)
63. D.V. Nicolau Jr., K. Burrage, R.G. Parton, et al., Identifying optimal lipid raft characteristics required to promote nanoscale protein-protein interactions on the plasma membrane. Mol. Cell. Biol. **26**(1), 313–323 (2006)
64. D.V. Nicolau Jr., J.F. Hancock, K. Burrage, Sources of anomalous diffusion on cell membranes: A Monte Carlo study. Biophys. J. **92**, 1975–1987 (2007)
65. B. Drawert, M.J. Lawson, L. Petzold, M. Khammash, The diffusive finite state projection algorithm for efficient simulation of the stochastic reaction-diffusion master equation. J. Chem. Phys. **132**(074101), 2010 (2010). doi:10.1063/1.3310809
66. T.T. Marquez-Lago, K. Burrage, Binomial tau-leap spatial stochastic simulation algorithm for applications in chemical kinetics. J. Chem. Phys. **127**, 104101 (2007)
67. A.B. Stundzia, C.J. Lumsden, Stochastic simulation of coupled reaction-diffusion processes. J Comp Phys **127**, 196–207 (1996)
68. F. Baras, M. Malek Mansour, Reaction-diffusion master equation: A comparison with microscopic simulations. Phys. Rev. E **54**(6), 6139–6148 (1996)
69. J. Elf, A. Doncic, M. Ehrenberg, Mesoscopic reaction-diffusion in intracellular signaling. Proc. SPIE **5110**, 114–125 (2003)
70. M. Ander, P. Beltrao, B. Di Ventura, et al., SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: Analysis of simple networks. Syst. Biol. **1**, 129–138 (2004)
71. R. Erban, S.J. Chapman, P.K. Maini, *A practical guide to stochastic simulations of reaction–diffusion processes*, arXiv:0704.1908 (2007
72. J. Elf, M. Ehrenberg, Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases. Syst. Biol. **1**, 230–236 (2004)
73. J. Hattne, D. Fange, J. Elf, Stochastic reaction-diffusion simulation with MesoRD. Bioinformatics **21**, 2923–2924 (2005)
74. S. Engblom, L. Ferm, A. Hellander, P. Loetstedt, Simulation of stochastic reaction-diffusion processes on unstructured meshes. SIAM J. Sci. Comput. **31**, 1774–1797 (2009)
75. R. Metzler, J. Klafter, The random walk's guide to anomalous diffusion: A fractional dynamics approach. Phys Reports **339**, 1–77 (2000)
76. D.S. Martin, M.B. Forstner, J.A. Kas, Apparent subdiffusion inherent to single particle tracking. Biophys. J. **83**(4), 2109–2117 (2002)
77. P.R. Smith et al., Anomalous diffusion of major histocompatibility complex class I molecules on HeLa cells determined by single particle tracking. Biophys. J. **76**(6), 3331–3344 (1999)
78. T.M. Jovin, W.L. Vaz, Rotational and translational diffusion in membranes measured by fluorescence and phosphorescence methods. Methods Enzymol. **172**, 471–513 (1989)
79. M. Weiss, H. Hashimoto, T. Nilsson, Anomalous protein diffusion in living cells as seen by fluorescence correlation spectroscopy. Biophys. J. **84**(6), 4043–4052 (2003)
80. B. Leitinger, N. Hogg, The involvement of lipid rafts in the regulation of integrin function. J. Cell Sci. **115**(Pt 5), 963–972 (2002)
81. H. Berry, Monte Carlo simulations of enzyme reactions in two dimensions: Fractal kinetics and spatial segregation. Biophys. J. **83**(4), 1891–1901 (2002)
82. R.A. Kerr et al., Fast Monte Carlo simulation methods for biological reaction-diffusion Systems in Solution and on surfaces. SIAM J. Sci. Comput. **30**(6), 3126 (2008)
83. R. Hilfer, L. Anton, Fractional master equations and fractal time random walks. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics **51**(2), R848–R851 (1995)

84. D. Fulger, E. Scalas, G. Germano, Monte Carlo simulation of uncoupled continuous-time random walks yielding a stochastic solution of the space-time fractional diffusion equation. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **77**(2 Pt 1), 021122 (2008)
85. F. Bezanilla, The voltage sensor in voltage-dependent ion channels. Physiol. Rev. **80**, 555 (2000)
86. B. Hille, *Ionic Channels of Excitable Membranes*, 3rd edn. (Sinauer Associates, Sunderland, MA, 1991). isbn:0878933239
87. B. Sakmann, E. Neher, *Single-Channel Recording* (Plenum, New York, 1995)
88. J.A. White, J.T. Rubinstein, A.R. Kay, Channel noise in neurons. Trends Neurosci. **23**, 131 (2000.) ISSN 0166-2236
89. E. Pueyo, A. Corrias, L. Virag, N. Jost, T. Szel, A. Varro, N. Szentandrassy, P.P. Nanasi, K. Burrage, B. Rodrıguez, A multiscale investigation of repolarization variability and its role in cardiac arrhythmogenesis. Biophys. J. **12**, 2892 (2011)
90. M. Lemay, E. de Lange, J.P. Kucera, Effects of stochastic channel gating and distribution on the cardiac action potential. J. Theor. Biol. **281**, 84 (2011.) ISSN 1095-8541
91. G. De Vries, A. Sherman, Channel sharing in pancreatic beta-cells revisited: Enhancement of emergent bursting by noise. J. Theor. Biol. **207**, 513 (2000)
92. J.R. Clay, L.J. DeFelice, Relationship between membrane excitability and single channel open-close kinetics. Biophys. J. **42**, 151 (1983). doi:10.1016/S0006-3495(83)84381-1
93. E. Schneidman, B. Freedman, I. Segev, Ion channel stochasticity may be critical in determining the reliability and precision of spike timing. Neural Comput. **10**, 1679 (1998.) ISSN 0899-7667, URL http://neco.mitpress.org/cgi/content/abstract/10/7/1679
94. P. Oosterhoff, A. Oros, M.A. Vos, Anadolu kardiyoloji dergisi. Anatolian journal of cardiology **7**(Suppl 1), 73 (2007.) ISSN 1302-8723, URL http://view.ncbi.nlm.nih.gov/pubmed/17584687
95. H. Mino, J.T. Rubinstein, J.A. White, Comparison of algorithms for the simulation of action potentials with stochastic sodium channels. Ann. Biomed. Eng. **30**, 578 (2002.) ISSN 0090-6964
96. R. Fox, Stochastic versions of the Hodgkin-Huxley equations. Biophys. J. **72**, 2068 (1997.) ISSN 00063495
97. X.J. Sun, J.Z. Lei, M. Perc, Q.S. Lu, S.J. Lv, Effects of channel noise on firing coherence of small-world Hodgkin-Huxley neuronal networks. The European Physical Journal B - Condensed Matter and Complex Systems **79**, 61 (2011). doi:10.1140/epjb/e2010-10031-3
98. R.F. Fox, Y.N. Lu, Emergent collective behavior in large numbers of globally coupled independently stochastic ion channels. Phys. Rev. E **49**, 3421 (1994)
99. I.C. Bruce, Evaluation of stochastic differential equation approximation of ion channel gating models. Ann. Biomed. Eng. **37**, 824 (2009.) ISSN 1521-6047
100. B. Sengupta, S.B. Laughlin, J.E. Niven, Comparison of Langevin and Markov channel noise models for neuronal signal generation. Phys. Rev. E **81**, 011918 (2010.), ISSN 1550-2376)
101. J.H. Goldwyn, N.S. Imennov, M. Famulare, E. Shea-Brown, Stochastic differential equation models for ion channel noise in Hodgkin-Huxley neurons. Phys. Rev. E **83**, 041908 (2011)
102. A.L. Hodgkin, A.F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. **117**, 500 (1952.) ISSN 0092-8240
103. T. Jahnke, W. Huisinga, Solving the chemical master equation for monomolecular reaction systems analytically. J. Math. Biol. **54**, 1–26 (2007.) ISSN 0303-6812
104. J.H. Goldwyn, E. Shea-Brown, The what and where of adding channel noise to the Hodgkin-Huxley equations. PLoS Comput. Biol. **7**, e1002247+ (2011). doi:10.1371/journal.pcbi.1002247
105. C.E. Dangerfield, D. Kay, K. Burrage, Stochastic models and simulation of Ion Channel dynamics. Procedia Computer Science **1**, 1581 (2010.) ISSN 18770509
106. C.E. Dangerfield, D. Kay, S. MacNamara, K. Burrage, A boundary preserving numerical algorithm for the Wright-fisher model with mutation. BIT Numer. Math. **52**, 283 (2011)
107. R. Lord, R. Koekkoek, D.V. Dijk, A comparison of biased simulation schemes for stochastic volatility models. Quantitative Finance **10**, 177 (2010)

108. C.E. Dangerfield, D. Kay, K. Burrage, Modelling ion channel dynamics through reflected stochastic differential, equations. Phys. Rev. E **85**, 051907 (2012)
109. D. Schnoerr, G. Sanguinetti, R. Grima, The complex chemical Langevin equation. J.Chem.Phys. **141**(2), 024103 (2014)
110. E. Schmidt, Über eine Klasse linearer funktionaler Differentialgleichungen. Math. Ann. **70**(4), 499–524 (1911)
111. F.R. Sharpe, A.J. Lotka, Contribution to the analysis of malaria epidemiology. IV. Incubation lag. Am. J. Epidemiology. **3**(supp1), 96–112 (1923)
112. V. Volterra, Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Memorie del R. Comitato Talassografico Italiano **43**, 1–142 (1927)
113. Y. Kuang, *Delay Differential Equations: With Applications in Population Dynamics* (Academic Press, Boston, MA, 1993)
114. J. Arino, P. van den Driessche, Time delays in epidemic models: Modeling and numerical considerations, in *Delay Differential Equations and Applications*, ed. by O. Arino et al. (Springer, New York, 2006), pp. 539–578
115. J. Lewis, Autoinhibition with transcriptional delay: A simple mechanism for the zebrafish somitogenesis oscillator. Curr. Biol. **13**(16), 1398–1408 (2003)
116. J. Srividhya, M.S. Gopinathan, A simple time delay model for eukaryotic cell cycle. J. Theor. Biol. **241**(3), 617–627 (2006)
117. S. Yi, P.W. Nelson, A.G. Ulsoy, *Time-Delay Systems: Analysis and Control Using the Lambert W Function* (New Jersey World Scientific, 2010)
118. M. Villasana, A. Radunskaya, A delay differential equation model for tumor growth. J. Math. Biol. **47**(3), 270–294 (2003)
119. M.V. Barbarossa, C. Kuttler, J. Zinsl, Delay equations modeling the effects of phase-specific drugs and immunotherapy on proliferating tumor cells. Math. Biosc. Eng **9**(2), 241 (2012)
120. K. Cooke, Y. Kuang, B. Li, Analyses of an antiviral immune response model with time delays. Canad. Appl. Math. Quart. **6**(4), 321–354 (1998)
121. C.-S. Kim, J.M. Ansermino, J.-O. Hahn, A comparative data-based Modeling study on respiratory CO_2 gas exchange during mechanical ventilation. Front. Bioeng. Biotechnol. **4**, 8 (2016). doi:10.3389/fbioe.2016.00008
122. D. Bratsun, D. Volfson, J. Hasty, L. Tsimring, Delay-induced stochastic oscillations in gene regulation. PNAS **102**(41), 14593–14598 (2005). doi:10.1073/pnas.0503858102
123. M.R. Roussel, R. Zhu, Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. Phys. Biol. **3**(4), 274–284 (2006)
124. X. Cai, Exact stochastic simulation of coupled chemical reactions with delays. J. Chem. Phys. **126**(12), 124108 (2007). doi:10.1063/1.2710253
125. D.F. Anderson, A modified next reaction method for simulating chemical systems with time dependent propensities and delays. J. Chem. Phys. **127**, 214107 (2007)
126. A. Leier, T.T. Marquez-Lago, Delay chemical master equation: Direct and closed-form solutions. Proc. R. Soc. A **471**, 20150049 (2015)
127. V. Sunkara, M. Hegland, An optimal finite state projection method. Procedia Comput. Sci. **1**, 1579–1586 (2010)
128. V. Wolf, R. Goel, M. Mateescu, T. Henzinger, Solving the chemical master equation using sliding windows. BMC Syst. Biol. **4**, 42 (2010)
129. K. Burrage, M. Hegland, S. MacNamara, R. Sidje, in *A Krylov-based finite state projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems*. Markov Anniversary Meeting: An International Conference to Celebrate the 150th Anniversary of the Birth of A.A. Markov (2006), pp. 21–38
130. B. Munsky, M. Khammash, The finite state projection algorithm for the solution of the chemical master equation. J. Chem. Phys. **124**(4), 044104 (2006)
131. R.B. Sidje, H.D. Vo, Solving the chemical master equation by a fast adaptive finite state projection based on the stochastic simulation algorithm. Math. Biosci. **269**, 10–16 (2015)
132. K.N. Dinh, R.B. Sidje, Understanding the finite state projection and related methods for solving the chemical master equation. Phys. Biol. **13**(3), 035003 (2016)

133. A. Borghans, R.J. de Boer, L.A. Segel, Extending the quasi-steady state approximation by changing variables. Bull. Math. Biol. **58**(1), 43 (1996)
134. E.A. Mastny, E.L. Haseltine, J.B. Rawlings, Two classes of quasi-steady-state model reductions for stochastic kinetics. J. Chem. Phys. **127**(9), 094106 (2007)
135. P. Thomas, A.V. Straube, R. Grima, The slow-scale linear noise approximation: An accurate, reduced stochastic description of biochemical networks under timescale separation conditions. BMC Syst. Biol. **6**(1), 39 (2012)
136. P. Thomas, R. Grima, A.V. Straube, Rigorous elimination of fast stochastic variables from the linear noise approximation using projection operators. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **86**(4 Pt 1), 041110 (2012)
137. M. Barrio, A. Leier, T.T. Marquez-Lago, Reduction of chemical reaction networks through delay distributions. J. Chem. Phys. **138**, 104114 (2013). doi:10.1063/1.4793982
138. A. Leier, M. Barrio, T.T. Marquez-Lago, Exact model reduction with delays: Closed-form distributions and extensions to fully bi-directional monomolecular reactions. J R Soc Interface (2014). https://doi.org/10.1098/rsif.2014.0108
139. A. Robson, K. Burrage, M.C. Leake, Inferring diffusion in single live cells at the single-molecule level. Phil Trans Roy Soc B **368**, 20120029 (2013)
140. G. Lillacci, M. Khammash, Parameter estimation and model selection in computational biology. PLoS Comput. Biol. **6**, e1000696 (2010)
141. T. Toni, D. Welch, N. Strelkowa, A. Ipsen, M.P.H. Stumpf, Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J Roy Soc Interface **6**, 187–202 (2009)
142. D.J. Wilkinson, Bayesian methods in bioinformatics and computational systems biology. Brief. Bioinform. **8**, 109–116 (2007)

Part IV
Diffusion Processes and Stochastic Modeling

Recent Mathematical Models of Axonal Transport

Chuan Xue and Gregory Jameson

1 Introduction

A neuron is a highly polarized cell that typically consists of a cell body, multiple dendrites, and a single axon (Fig. 1). Both dendrites and axons are thin structures that extend away from the cell body. Dendrites usually branch multiple times, forming a dendritic tree. Axons are often much longer than dendrites and have a more-or-less regular shape. A neuron receives electrical signals from other cells through dendrites and propagates the signals to other cells through its axon. The electrical properties of a neuron are critically dependent on its shape.

An axon in the peripheral nervous system has a cross-sectional diameter of 1–10 μm but can be as long as 1 m in humans. The ability for an axon to survive and maintain its shape largely depends on a dynamic system of intracellular polymers, called the *axonal cytoskeleton*, and the active movement of various organelles and macromolecular proteins along the cytoskeleton, known as *axonal transport*.

The axonal cytoskeleton includes microtubules, microfilaments, and neurofilaments. Microtubules are long, hollow, polarized polymers with diameter 25 nm and persistence length $80 \pm 20\,\mu$m [1]. They align axially along the axon, with plus ends pointing away from the cell body. They do not extend over the whole length of an axon but rather form an overlapping array from the cell body to the axon terminal [2, 3]. One of their primary purposes is to serve as tracks for the long-range transport of membranous organelles and macromolecular proteins [4, 5]. The transport is

C. Xue (✉)
Department of Mathematics, Ohio State University, Columbus, OH, USA
e-mail: cxue@math.osu.edu

G. Jameson
Biophysics Graduate Program, Ohio State University, Columbus, OH, USA
e-mail: jameson.61@osu.edu

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_12

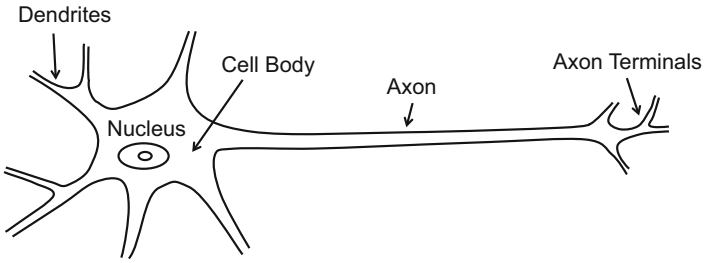


Fig. 1 Schematic drawing of a typical unmyelinated neuron to show its overall shape

powered by the molecular motors kinesin and dynein, which move cargoes towards the plus and minus ends of microtubules, respectively [6].

Microfilaments (actin) are much shorter polarized filaments with diameter 5–7 nm and persistence length about 18 μm [7, 8]. They are particularly abundant beneath the axon membrane, forming evenly spaced ring-like structures that wrap around the circumference of the axon along axonal shafts [9]. Microfilaments are also present near microtubules and enriched in growth cones and axon terminals [10]. They are involved in the short-range, lateral transport of organelles and proteins, powered by the molecular motor myosin [11].

Neurofilaments are the intermediate filaments of neurons. They are long, flexible, non-polarized polymers with diameter about 10 nm and persistence length about 200–450 nm [12, 13]. They are brush-like polymers with densely packed sidearms extending away from the polymer backbone. In contrast to microtubules and microfilaments, they do not serve as tracks for the transport of cargoes. Rather, they function as space-filling structures to increase the axonal caliber [14] and occupy most of the axonal volume in large axons [15]. Their sidearms help maximize their space-filling properties.

Axonal transport has been classified into two categories, fast axonal transport and slow axonal transport, depending on the average movement rate of the cargoes [5]. Fast axonal transport is the transport of membranous organelles such as mitochondria and lysosomes, either unidirectionally or bidirectionally, at rates of up to hundreds of millimeters per day. Slow axonal transport is the transport of cytoskeletal polymers and cytosolic proteins, including neurofilaments, with much slower rates of a few millimeters per day. The fundamental difference between fast and slow axonal transport is not the instantaneous velocity of the cargoes, but the frequency and persistency of the movement. For example, neurofilaments undergo infrequent bidirectional transport along microtubules with short bouts of movements interrupted by prolonged pauses and spend more than 90% of their time pausing [16, 17].

Axonal transport is disrupted in many neurodegenerative diseases including Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis (also known

as Lou Gehrig's disease), Huntington's disease, hereditary spastic paraplegia, and Charcot-Marie-Tooth disease [18–20]. In these diseases, aberrant accumulations of certain cellular components and excessive focal swelling of the axon were observed, which eventually lead to axon degeneration. While some of these diseases are caused by direct mutations of the molecular motors or adaptor proteins that interact with motors [7, 21–25], the underlying mechanisms of many others are not clear. Intriguingly, some of these disorders are associated with cross-sectional segregation of the axonal microtubules and neurofilaments, which is never observed in normal axons [26–31].

Some basic questions that arise in studying axonal transport, either from experimental or mathematical standpoints, are as follows:

- How do organelles and proteins move to their destination in normal situations? What perturbations of axonal transport can give rise to local cargo pileups and axonal swellings as observed in diseases?
- How is the mixed distribution of microtubules and neurofilaments established and maintained in normal axons? What lead to their cross-sectional separation in toxic neuropathies? How is the cytoskeleton segregation related to subsequent axonal swelling?
- Large myelinated axons demonstrate a spatially periodic structure with naturally occurring narrowing points called *nodes of Ranvier*, with roughly the same number of microtubules but much fewer neurofilaments and higher organelle density [32]. How is axonal transport regulated at the nodal regions? How is the differential distribution of microtubules, neurofilaments, and organelles established near and away from nodes of Ranvier?

Due to the complicated biology of axonal transport, mathematical modeling is an essential tool to address the above questions. In this chapter, we summarize recent mathematical models that have advanced our understanding of these questions. In Sect. 2, we review earlier models that describe cargo movement in normal axons as a 1D process. In Sect. 3, we reviewed and extended recent modeling effort that describes the cross-sectional dynamics of microtubules and associated cargoes, in particular, the microtubule–neurofilament segregation phenomena observed in neuropathies. In Sect. 4, we describe a few open problems in modeling axonal transport.

The molecular basis of axonal transport is not fundamentally different from intracellular transport in other cellular systems. Our knowledge of molecular motors and intracellular transport in the general setting has been expanding consistently over the last few decades, especially regarding how a single motor or a group of motors coordinate with each other to transport a cargo in in vitro experiments. This has resulted in a large number of mathematical models on these topics, and we refer interested readers to [33, 34] for reviews.

2 Models of Axonal Transport in Healthy Axons

Pulse-labeling methods have been used to study axonal transport since the 1950s [5, 35]. In these experiments, radio-labeled amino acids are injected into the vicinity of nerve cell bodies, which will then be taken up by the cell bodies, incorporated into proteins, and transported to axons in association with various cargoes. The labeled amino acids are usually injected for a few hours or continuously, and their intensities are measured at different locations of the axon at later times. If the injection occurs only for a few hours, the radioactivity forms a Gaussian wave that moves towards the axon terminal and spreads out over time. If the injection occurs continuously, the radioactivity forms a wave front that propagates into the axon at a constant speed.

Motivated by these experiments, different hypotheses on the molecular mechanisms of fast and slow axonal transport have been formulated in the 1980s, and mathematical models have been used to test these hypotheses [36–40]. A PDE model for fast axonal transport was constructed in [38], which accounted for the reversible binding of organelles, kinesins, and microtubules, as well as the movement of organelles along microtubules engaged through motor proteins. The model qualitatively matches the moving waves observed in the pulse-labeling experiments. A PDE model for the less-understood slow axonal transport of neurofilaments was developed in [40], based on a hypothetical unidirectional transport engine with a constant velocity that slow cargoes either bind directly or piggy-back through other structures in a reversible manner.

A common assumption of these models is that the movement of cargoes is independent and homogeneous in 1D. The governing equations are of reaction-hyperbolic type with the following general form:

$$\partial_t p_i(x,t) + v_i \partial_x p_i(x,t) = f_i(p_1, \ldots, p_N), \quad t > 0, \quad 1 \le i \le N. \tag{1}$$

Here $p_i(x,t)$ is the density of the i-th species, which is either a cargo with velocity v_i or other proteins involved in the transport process with velocity $v_i = 0$. The functions $f_i(p_1, \ldots, p_N)$ enclose chemical reactions among cargoes, motors, and auxiliary proteins. Simulations suggest that this class of models, with either linear or nonlinear reaction terms, admit the so-called approximate traveling wave solutions, meaning the actively moving species have wave-like profiles that propagate along the x-axis with slowly diminishing magnitudes [41]. This corresponds to the wave-like propagation of radioactivity observed in pulse-labeling experiments.

Considerable progress has been made in the past 15 years in uncovering the molecular mechanisms of slow axonal transport due to the advancement of fluorescence microscopy. It is now understood that slow axonal transport of neurofilaments also depends on the molecular motors kinesin and dynein, and move bidirectionally in a rapid and intermittent manner [16, 42]. Mathematical models have been developed accordingly for the bidirectional movement of neurofilaments and their transitions among several different velocity states [43–46]. The governing equations, essentially a linearized version of (1), are

$$\partial_t p_i(x,t) + v_i \partial_x p_i(x,t) = \frac{1}{\varepsilon} \sum_{i,j=1}^{N} k_{ij} p_j(x,t), \quad t > 0, \quad N \leq 2. \tag{2}$$

Here $p_i(x,t)$ is the (probability) density of neurofilaments with velocity v_i, k_{ij} are the transition rates between different velocity states, and the small parameter ε illustrates that the state transitions occur on a much shorter time scale than the relocation of the cargo population. The transition rates k_{ij} in these models have also been extracted from experimental data [47, 48]. We note that these models distinguish neurofilaments by the phenomenological velocity states instead of their association with motors and microtubules.

Using formal perturbation methods, Reed et al. [41] demonstrated the existence of approximate traveling wave solutions of (1) with only one moving species and linear reaction terms, which essentially is the same as (2) with one nonzero v_i. Friedman and Craciun [49] proved this result rigorously for the general model (2) with arbitrary N. Such analyses have also been extended to account for cargo diffusion in [50–52].

Motivated by PDE models of axonal transport, probabilistic methods have been developed for linear reaction-hyperbolic models of axonal transport and their spatially discretized forms [53, 54]. Stochastic models have also been developed for intracellular transport in dendrites, mainly focusing on the delivery of cargoes to synapses in the dendritic tree [55–58].

3 Models of Axonal Transport in Neurological Diseases

In healthy axons, microtubules and neurofilaments align along the axon and are interspersed in axonal cross-sections [2, 59, 60]. However, in many toxic neuropathies these two populations of polymers separate radially, with microtubules and organelles located near the long axis of the axon and neurofilaments displaced to the periphery near the axonal membrane (Fig. 2). This striking cytoskeletal reorganization proceeds to axonal swelling and has been reported in neurodegenerative disorders including giant axonal neuropathy and Charcot-Marie-Tooth disease, in neurotoxic neuropathies induced by exposure to a range of neurotoxins, and in a transgenic mouse expressing a mutant neurofilament protein [26–31, 61–75].

The microtubule–neurofilament segregation phenomenon has been studied most extensively in axons treated by the toxin 3,3′-iminodiproprionitrile (IDPN) and 2,5-hexanedione (Fig. 2). IDPN is closely related to the food poison 3-aminopropionitrile that causes the neurological disorder lathyrism [76–79], and 2,5-hexanedione is a metabolite of the industrial solvent hexane. Systematic administration of IDPN to rats causes microtubules and neurofilaments to segregate within a few hours, and leads to excessive focal accumulations of neurofilaments and

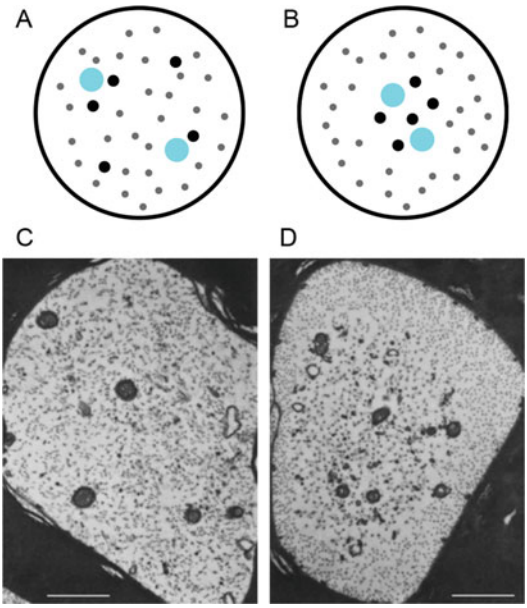


Fig. 2 Radial separation of microtubules and neurofilaments in experiments. (**a**) Schematic drawing that illustrates the normal distribution of microtubules (*large black dots*), neurofilaments (*small grey dots*), and organelles (*cyan disks*) in untreated axons. (**b**) Schematic drawing that illustrates the segregated components in IDPN-treated axons. (**c**) Electron micrograph of a normal axonal cross-section, with microtubules, neurofilaments, and organelles mixed together. *Big dots*: microtubules; *small dots*: neurofilaments; objects: organelles. (**d**) Electron micrograph of an IDPN-treated axon, with microtubules and organelles located in the center and neurofilaments migrated to the periphery. The *black region* outside of the axon is the myelin sheath. Scale bars: 1 μm. Reproduced from [61]

axonal swelling in a few days [61, 80, 81]. If IDPN is washed out, the organization of microtubules and neurofilaments reverses to normal [61]. These phenomena have been observed for over 30 years now, but the underlying mechanisms are still largely unexplored.

Pulse-labeling experiments showed that IDPN selectively impairs the longitudinal transport of neurofilaments but does not affect the transport of organelles [82]. How impairment of neurofilament transport in the longitudinal direction relates to the radial segregation of neurofilaments from microtubules is unclear. It is attractive to postulate extra mechanisms to explain the polymer separation, because it is clear that if neurofilaments are separated from microtubules they cannot be transported along microtubules. However, recent mathematical modeling suggested that the polymer separation can be explained as a consequence of the impairment of neurofilament transport instead and extra mechanisms are not necessary [83].

3.1 Radial Cytoskeleton Separation Explained by a Stochastic Model

In [83], a stochastic model was developed to track the distribution of microtubules, neurofilaments, and organelles in a cross-section of an axon. The model describes microtubules, neurofilaments, and organelles as nondeformable particles with center positions $\mathbf{x}_i^k$ in a 2D circular domain. Here $k = M, N$, or O is the index for particle type: M for microtubule, N for neurofilament, O for organelle; and i with $0 \le i \le n^k$ is the index for the k-type particle where n^k is the total number of k-type particles.

These particles interact with each other through three key mechanisms. The first is the slow axonal transport of neurofilaments. This is included in the model by allowing random binding and unbinding of a neurofilament and a microtubule within a binding radius, random arrival of a neurofilament near a microtubule, and random departure of a neurofilament that is bound to a microtubule. The binding radius reflects the length of the molecular motors, kinesin or dynein, that bridge a neurofilament and a microtubule. The directionality of neurofilament transport, either anterograde or retrograde, is not distinguished. The force between an engaged neurofilament–microtubule pair is assumed to be a linear elastic spring force, with magnitude $S = \kappa^{MN} d$, where d is the surface distance and κ^{MN} is the spring constant of the molecular motor complex connecting a microtubule and a neurofilament.

The second mechanism is the fast axonal transport of organelles. This is included in a similar way as neurofilament transport. The difference is that when an organelle enters the domain, it moves persistently without stopping and as a result its cross-sectional radius first increases from 0 to maximum and then decreases to 0. This relaxation process mimics the change of radius of the organelle inside an axonal cross-section and prevents overlapping particles in computation.

The third mechanism is volume exclusion. This is incorporated through pairwise volume exclusive forces that take into account the biophysical properties of the particles. The magnitude of the repulsion forces is assumed to be

$$R = \varepsilon_r (L_r/d - 1)_+, \tag{3}$$

where ε_r is a force prefactor, L_r is the maximum repulsion distance, d is the surface distance between the particle pair, and the subscript $+$ means taking the positive part of the function.

By neglecting inertia, the particle positions are governed by a system of SDEs

$$\mathrm{d}\mathbf{x}_i^k = \mathbf{F}_i^k/\mu^k \,\mathrm{dt} + \sigma_k \,\mathrm{d}\mathbf{W}_i^k, \quad 1 \le i \le n^k,\ k = M, N, O. \tag{4}$$

The force $\mathbf{F}_i^k$ is the total force acting on the i-th particle of type k by all other particles specified in the three mechanisms. Because particle binding and unbinding are treated as first-order stochastic processes, $\mathbf{F}_i^k$ is a stochastic force. Denote the pairwise repulsion acting on the i-th particle of type k by the j-th particle of type l by $\mathbf{R}_{ij}^{kl}$, and similarly the spring force by $\mathbf{S}_{ij}^{kl}$. Denote the binding state of two particles

by s_{ij}^{kl}, which jumps between 0 and 1 if one of the particles is a microtubule and the other is a cargo. With these notations we have

$$\mathbf{F}_i^M = \sum_{1\le j\le n_M} \mathbf{R}_{ij}^{MM} + \sum_{1\le j\le n_N} \left(\mathbf{R}_{ij}^{MN} + s_{ij}^{MN}\mathbf{S}_{ij}^{MN}\right) + \sum_{1\le j\le n_O} \left(\mathbf{R}_{ij}^{MO} + s_{ij}^{MO}\mathbf{S}_{ij}^{MO}\right), \quad 1 \le i \le n^M,$$

$$\mathbf{F}_i^N = \sum_{1\le j\le n_M} \left(\mathbf{R}_{ij}^{NM} + s_{ij}^{NM}\mathbf{S}_{ij}^{NM}\right) + \sum_{1\le j\le n_N} \mathbf{R}_{ij}^{NN} + \sum_{1\le j\le n_O} \mathbf{R}_{ij}^{NO}, \quad 1 \le i \le n^N,$$

$$\mathbf{F}_i^O = \sum_{1\le j\le n_M} \left(\mathbf{R}_{ij}^{OM} + s_{ij}^{OM}\mathbf{S}_{ij}^{OM}\right) + \sum_{1\le j\le n_N} \mathbf{R}_{ij}^{ON} + \sum_{1\le j\le n_O} \mathbf{R}_{ij}^{OO}, \quad 1 \le i \le n^O,$$

The constant μ^k in (4) denotes the drag coefficient of the k-type particle, $\mathbf{W}_i^k$ are independent 2D Wiener processes modeling the random motion of these particles, and σ_k gives the amplitude of the Brownian motion. Because the number of particles in the domain is stochastic due to cargo arrival and departure, the number of equations in the model is stochastic as well. All model parameters are physically meaningful and have been estimated using extensive experimental data.

The model provided the first mechanistic explanation for the cytoskeleton segregation phenomena shown in Fig. 2. Simulations of the model demonstrate that organelles can randomly bind with multiple microtubules simultaneously as they move along the axon. This provides a means of indirect interaction among microtubules and gradually cause microtubules to move closer to each other. We call this the "clustering effect" of organelle transport. On the contrary, neurofilaments mainly bind with a single microtubule, frequently unbind, and disperse microtubules apart from each other. We call this the "dispersing effect" of neurofilament transport. In the absence of neurofilament transport, organelles pull microtubules together to the center of the domain within hours, in a similar way as observed in the IDPN experiments.

If neurofilament transport is restored, microtubules remix with the neurofilaments (Fig. 3).

The clustering effect of moving organelles was further investigated by varying the maximum number of microtubules (M) that an organelle can bind to simultaneously. Initially, microtubules and neurofilaments are uniformly distributed in cross-section, and neurofilament transport is blocked but organelle transport is normal. Figure 4 plots the mean pairwise distance of microtubules averaged over five realizations, which provides a measure of how separated microtubules and neurofilaments are over time. The error bars indicate the standard deviation over all realizations for each situation. As M increases, the clustering effect becomes stronger and the average microtubule distance decreases faster. If an organelle is only allowed to bind to one or two microtubules concurrently, then organelles cannot

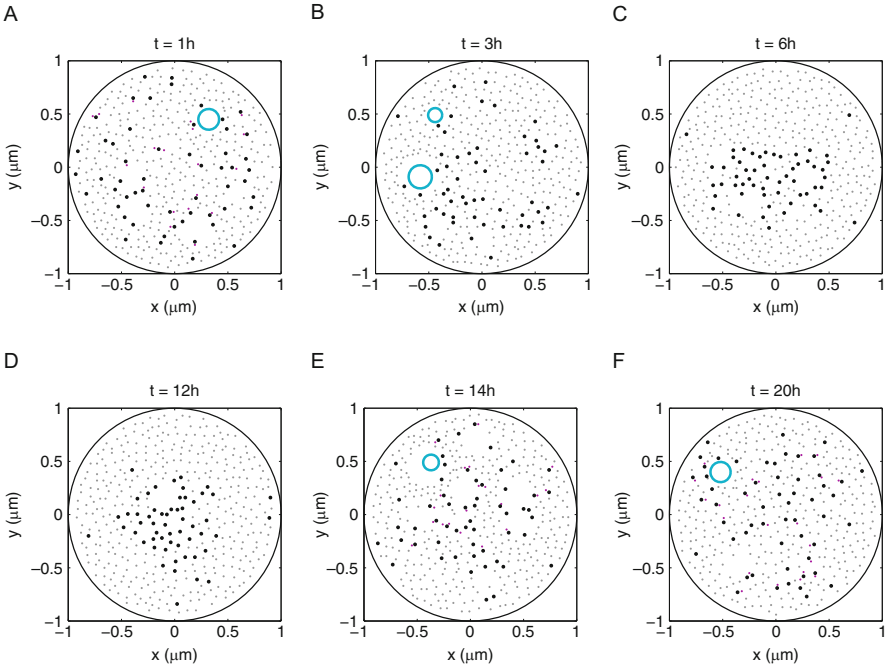


Fig. 3 A realization of the model in [83] shows reversible segregation in IDPN-treated axons. Neurofilament transport is blocked starting at $t = 1$ h and restored at $t = 13$ h. *All panels* are from a single realization. *Large black dots* are microtubules; *small grey dots* are free neurofilaments; *small purple dots* are neurofilaments engaged with microtubules; *large cyan circles* are organelles. Reproduced from [83] with permission. (**a**) Normal axon. (**b**) 2 h after IDPN. (**c**) 5 h after IDPN. (**d**) 11 h after IDPN. (**e**) 1 h after washout. (**f**) 7 h after washout

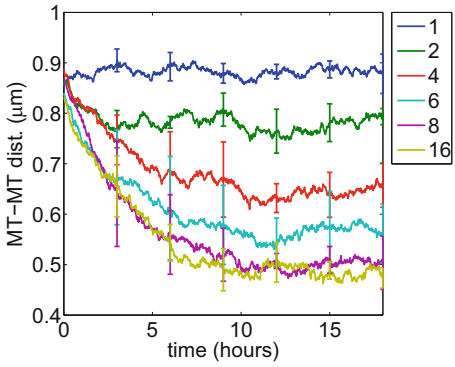


Fig. 4 The clustering effect of moving organelles. The mean of pairwise distance of microtubules is plotted over time. The maximum number of microtubules that an organelle can bind with simultaneously is set to be 1, 2, 4, 8, 16 for the *blue*, *green*, *red*, *cyan*, *purple*, and *yellow curves*, respectively. *Each curve* represents the average over five realizations with *error bars* indicating standard deviations. Reproduced from [83] with permission

effectively cluster microtubules even after 18 h (blue, green curves). This further confirms that the microtubule–neurofilament segregation does not occur naturally but is caused by the active zippering of passing organelles.

Moreover, the model makes several testable predictions. One of the predictions is that the extent and rate of microtubule–neurofilament segregation depends on the flux rate and size of the moving organelles: the more and the larger the organelles are, the faster the segregation occurs. This prediction could be tested by perturbing organelle transport inside an IDPN-treated axon. The model also suggests that the segregation occurs through the merging of small microtubule clusters, which has been previously observed in experiments [84].

Fast axonal transport and slow axonal transport have usually been studied as independent processes in previous experiments and mathematical models. However, the model in [83] highlights that fast and slow axonal transport are competing processes: different cargoes compete for microtubule tracks, and different balances of fast and slow axonal transport can lead to different cargo distributions.

3.2 Further Investigation of Microtubule Zippering by Moving Organelles

In [83], microtubule–microtubule interaction through moving organelles was incorporated in a highly simplified manner: microtubules and organelles were represented by disks that interact through elastic forces in the axonal cross-section. Microtubules are long polymers in axons—when an organelle moves along multiple microtubules simultaneously, it pulls them together, similar to a zipper. In this section, we develop a more detailed model to simulate this zippering mechanism, treating microtubules as rigid rods and organelles as spherical nondeformable particles. We restrict the movement of microtubules and organelles in 2D.

We consider the situation with one organelle moving along two microtubules (Fig. 5). We assume that the organelle has N kinesin motors bound on its surface at fixed, equidistant locations. Each motor can bind to and unbind from each microtubule with constant rates k_{on} and k_{off}, respectively. Binding can only occur if the motor head is within a capturing distance that is the natural length of the spring l_0, the binding location is randomly chosen so that the length of the motor is smaller than l_0. The motors are not allowed to intersect the organelle. Once bound, the motor head moves towards the plus end of the microtubule in $\delta = 8\,\text{nm}$ steps with rates $k_{\text{mov}} = v/\delta$, where $v = 1\,\mu\text{m/s}$ is the average speed of a motor. If the motor spring force is greater than $F_{\text{stall}} = 8\,\text{pN}$, then the motor is not allowed to move forward [85, 86]. We model microtubule-motor-organelle cross-bridges as extensional springs along the motor with magnitude $\kappa(l-l_0)_+$, where κ is the spring constant, l is the length of the motor spring in the stretched configuration, and the subscript "+" means taking the positive part. Microtubules and organelles interact

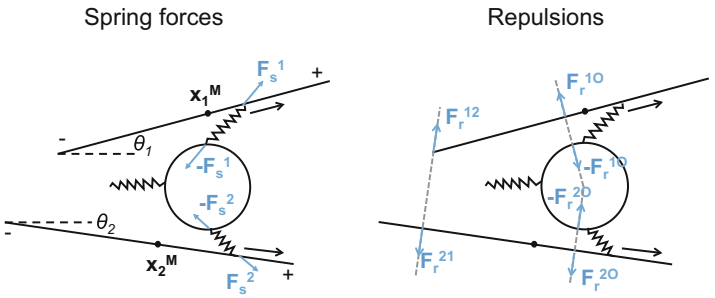


Fig. 5 Model schematics. *Left*: illustration of the model variables and the spring forces. *Right*: the repulsive forces. The *black arrows* indicate the direction of movement of the motor heads

through volume exclusion, modeled by repulsive forces acting at the nearest points of the particle pairs with magnitude given by (3).

Denote the center position of the organelle and the microtubules by $\mathbf{x}^O$ and $\mathbf{x}_i^M$, $i = 1, 2$, respectively, and their angles to the x-axis by θ^O and θ_i^M. The governing equations are

$$
\begin{aligned}
\mathrm{d}\mathbf{x}^O &= \Big(-\sum \mathbf{F}_s^1 - \sum \mathbf{F}_s^2 - \mathbf{F}_r^{1O} - \mathbf{F}_r^{2O} \Big) \Big/ \mu_t^O \,\mathrm{d}t + \sigma_t^O \,\mathrm{d}\mathbf{W}_\mathbf{t}^\mathbf{O}, \\
\mathrm{d}\theta^O &= \Big(-\sum \mathbf{l}_s^{O1} \times \mathbf{F}_s^1 - \sum \mathbf{l}_s^{O2} \times \mathbf{F}_s^2 - \mathbf{l}_r^{O1} \times \mathbf{F}_r^{1O} - \mathbf{l}_r^{O2} \times \mathbf{F}_r^{2O} \Big) \Big/ \mu_r^O \,\mathrm{d}t + \sigma_r^O \,\mathrm{d}W_r^O, \\
\mathrm{d}\mathbf{x}_i^M &= \Big(\sum \mathbf{F}_s^i + \mathbf{F}_r^{iO} + \mathbf{F}_r^{ij} \Big) \cdot \Big(\mathbf{e}_\parallel^i \otimes \mathbf{e}_\parallel^i \Big/ \mu_a^M + \mathbf{e}_\perp^i \otimes \mathbf{e}_\perp^i \Big/ \mu_t^M \Big) \,\mathrm{d}t \\
&\quad + \Big(\sigma_a^M \mathbf{e}_\parallel^i \otimes \mathbf{e}_\parallel^i + \sigma_t^M \mathbf{e}_\perp^i \otimes \mathbf{e}_\perp^i \Big) \cdot \mathrm{d}\mathbf{W}_t^M, \quad i, j = 1, 2, \quad i \neq j, \\
\mathrm{d}\theta_i^M &= \Big(\sum \mathbf{l}_s^{iO} \times \mathbf{F}_s^i + \mathbf{l}_r^{iO} \times \mathbf{F}_r^{iO} + \mathbf{l}_r^{ij} \times \mathbf{F}_r^{ij} \Big) / \mu_r^M \,\mathrm{d}t + \sigma_r^M \mathrm{d}W_r^M, \quad i, j = 1, 2, \quad i \neq j.
\end{aligned}
\tag{5}
$$

Here the forces are defined in Fig. 5; $\mathbf{l}_s^{Oi}$ and $\mathbf{l}_r^{Oi}$ are the vectors pointing from $\mathbf{x}_i^O$ to the positions of the forces acting on the organelle; $\mathbf{l}_s^{iO}$, $\mathbf{l}_r^{iO}$, and $\mathbf{l}_r^{ij}$ are the vectors from $\mathbf{x}_i^M$ to the positions of the forces acting on the microtubules; the sums are taken over all the molecular motors; $\mathbf{e}_\parallel^i$ and $\mathbf{e}_\perp^i$ are unit vectors in the parallel and perpendicular directions of Microtubule i, respectively; $\mathbf{W}_t^k$ and W_r^k with $k = O, M$ are independent Wiener processes; μ_k^l are drag coefficients and $\sigma_k^l = \sqrt{2k_BT/\mu_k^l}$ where k_B is the Boltzmann constant and T is the temperature assumed to be room temperature. The spring forces are stochastic due to the random binding and unbinding between motors and microtubules.

The model parameters are summarized in Table 1. The natural length of the motor spring l_0 is estimated to be the length of a kinesin motor [87]. The parameter ε_r is calculated using the following way. We assume that when an organelle is transported along a single microtubule, the repulsion acting on the organelle balances the spring

Table 1 Model parameter values

Parameter	Description	Values	Notes and references
l_0	Natural length of the motor spring	80 nm	[87]
v	Motor speed	1 μm/s	[5]
L_r	Characteristic repulsion distance	0.1212 μm	[83]
ε_r	Repulsion prefactor	0.268 pN	Estimated
κ	Spring constant	300 pN/μm	[88, 89]
μ_t^O	Translational drag coefficient for organelles	11.3 pN s/μm	Calculated
μ_r^O	Rotational drag coefficient for organelles	0.339 pN s μm	Calculated
μ_a^M	Translational drag coefficient for microtubules parallel with its axis	168 pN s/μm	[90, 91]
μ_t^M	Translational drag coefficient for microtubules perpendicular to its axis	296 pN s/μm	[90, 91]
μ_r^M	Rotational drag coefficient for microtubules perpendicular to its axis	6.17×10^4 pN s μm	Calculated
δ	Kinesin step size	8 nm	[92, 93]
k_{on}	Motor-microtubule binding rate	8/s	Estimated
k_{off}	Motor-microtubule unbinding rate	2/s	[86, 94]
k_{mov}	Motor stepping rate	125/s	$= v/\delta$
N	Number of motors on the organelle	8	Estimated

force and its viscous drag, which leads to $\frac{\varepsilon_r(L_r/d_0-1)}{d_0} = \frac{\kappa(l-l_0)}{l} = \frac{\mu_t^O v}{\sqrt{l^2-d_0^2}}$, where $d_0 = 17$ nm is the observed average surface distance between a microtubule and a cargo engaged on it [95]. Solving ε_r and l gives the parameter ε_r. The drag coefficient μ_t^O and μ_r^O are calculated using the formulas $\mu_t^O = 6\pi\eta r_0$ and $\mu_r^O = 8\pi\eta r_0^3$ [96], where we took $\eta = 4\text{pN s}/\mu\text{m}^2$ [97]. The translational drag coefficients μ_a^M and μ_t^M are calculated using the formulas derived in [90] and reviewed in [91, p. 347]. The rotational drag coefficient μ_r^M is calculated using the following method. Let a uniform rod of length L be placed horizontally, with the center of mass at the origin. Let the rod rotate around the origin with angular velocity ω. The torque of the drag on the rod is $\tau = \int_{-L/2}^{L/2} \mu_t^M/L \cdot \omega x^2 \, \mathrm{d}x = \mu_t^M \omega L^2/12$. Thus we have $\mu_r^M = \tau/\omega = \mu_t^M L^2/12$. The rate k_{on} was estimated so that the average number of motors on each microtubule that bear force equals 1.

3.2.1 Persistent Motor-Microtubule Binding and Motor Movement

We first considered the deterministic version of the model with a single motor persistently bound to each microtubule that moves at a constant speed v and no Brownian motion of all objects (Fig. 6). We assumed that the organelle radius is

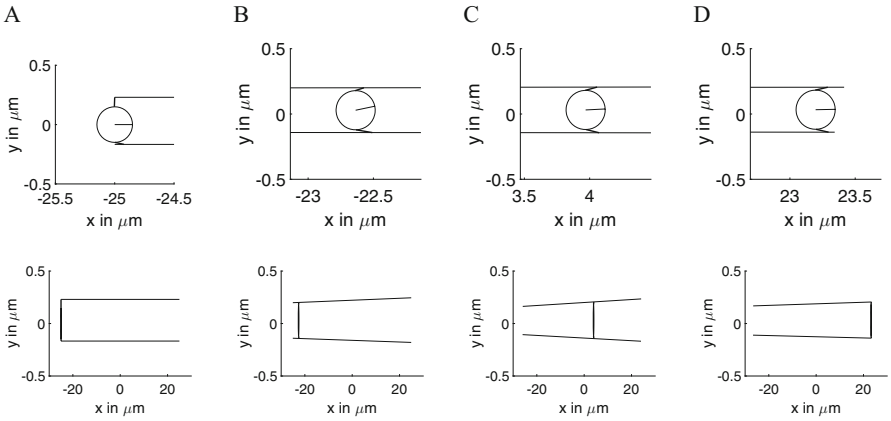


Fig. 6 Microtubule zippering predicted by the deterministic version of the stochastic model (5). The relative positions of the microtubules and the organelle at $t = 0$ s (**a**), 2.5 s (**b**), 30 s (**c**), 49.875 s (**d**). The *line* inside the organelle is included to illustrate the rotation of the organelle

$r_0 = 150$ nm and the microtubule length is 50 μm. We set the initial conditions to mimic the situation that the organelle had been moving along the bottom microtubule and started to interact with the top microtubule (Fig. 6a). At $t = 0$, the organelle has center position $\mathbf{x}^O = (-25\,\mu\text{m}, 0)$ and the two microtubules are placed in parallel to the x-axis with $x_1^M = x_2^M = 0$, $y_1^M = r_0 + l_0 = 230$ nm, and $y_1^M = -r_0 - d_0 = -167$ nm. The organelle is attached to both microtubules near their left ends, with the bottom microtubule bearing the full load and the corresponding motor oblique. The model was simulated using Forward Euler's Method with $\Delta t = 5$ ms. As the organelle moves to the right, both motors bear forces, and the microtubules move closer and become tilted due to the spring forces acting on them (Fig. 6b, c). The simulation was stopped once the right end passes the right end of a microtubule (Fig. 6d).

Figure 7 demonstrates the vertical motion of the center of each microtubule as well as their left and right ends. Both microtubule centers move significantly closer to the x-axis. The left ends move closer to each other before $t = 30$ s and space apart slightly afterwards as the organelle gets closer to their right ends. By the end of the simulation, the vertical distance between the left ends is smaller than the diameter of the organelle. The right ends move apart initially due to the rotation of the microtubules caused by the motors' pulling but move closer to each other as the organelle proceeds to the right.

3.2.2 Random Motor-Microtubule Binding and Motor Movement

We next simulated the full model (5) with the random binding, unbinding, and stepping of the molecular motors on microtubules, as well as Brownian motion of the microtubules and the organelle (Fig. 8). The initial setup of the simulation

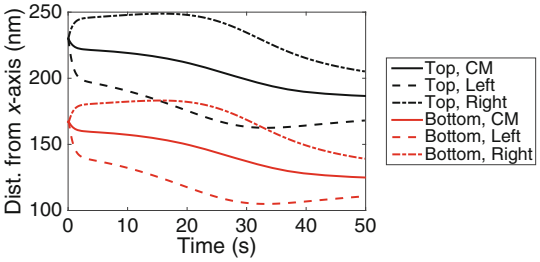


Fig. 7 The distance from the center and the ends of each microtubule to the x-axis over time. *Black curves*: the microtubule initially above the x-axis. *Red curves*: the microtubule initially below the x-axis. *Solid lines*: center of mass. *Dashed lines*: left end. *Dash-dotted lines*: right end

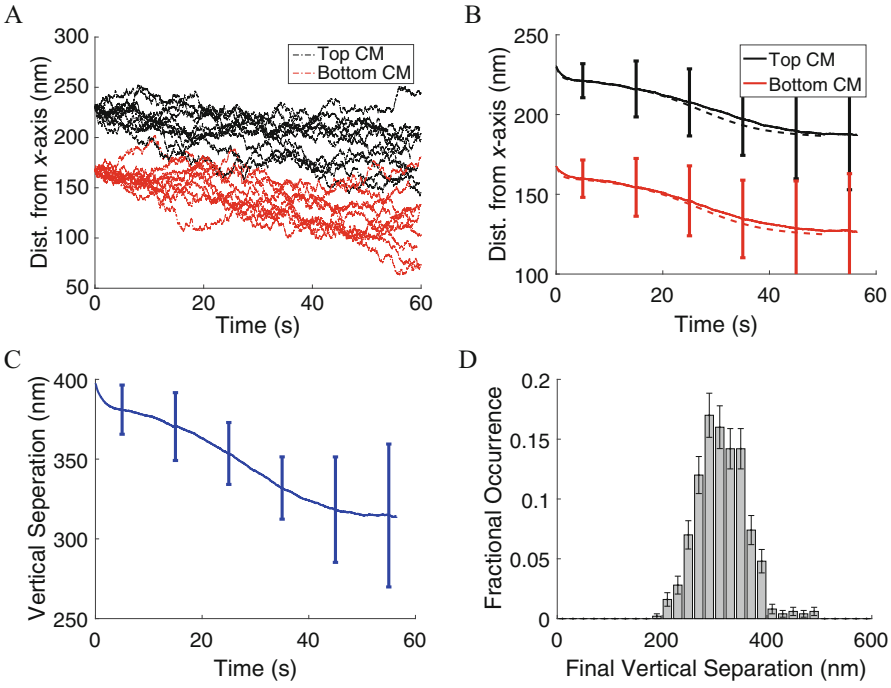


Fig. 8 Microtubule zippering predicted by the stochastic model (5) with random motor binding. (**a**) Microtubule center positions over time in ten realizations. (**b**) Mean and standard deviation of the microtubule center positions calculated using 500 realizations. The *dashed lines* are predictions from the deterministic version of the model. (**c**) Mean and standard deviation of the vertical separation of the microtubule centers. (**d**) Histogram of the final vertical separation of 500 realizations. The *error bars* indicate the statistical error in fractional occurrence

was the same as in Fig. 6, and the parameters used are included in Table 1. The model was simulated using the same algorithm as in [83] in MATLAB with random seed and $\Delta t = 1$ ms for 500 realizations. Figure 6 shows the evolution of the microtubule centers in the first ten realizations. Notable decrease in vertical

separation between the microtubules was observed in 484 realizations, and the average decrease rate agreed with the deterministic limit of the model quantitatively (Fig. 8b–d). These results confirm the zippering effect of moving organelles in the presence of stochastic noise. The model can be further refined by incorporating microtubules as flexible rods instead of stiff rods.

4 Open Problems

Axonal transport has been modeled using both continuous PDE approaches and stochastic individual-based approaches in 1D (see Sect. 2). These models assume that the axonal cargoes move independently as a homogeneous random walk. However, cargoes live in a crowded environment, and all cargoes and cytoskeletal polymers interact with each other stochastically via active motor-based transport and passive volume exclusion. Disruptions of particle interactions can lead to complicated phenomena such as axonal swelling and cytoskeleton segregation as observed in neurodegenerative diseases. To understand these phenomena, new mathematical models and methods must be developed to include particle interactions.

Recently progress has been made in integrating particle interactions to help understand cytoskeleton segregation in IDPN-treated axons (see Sect. 3.1). The stochastic particle-based model developed in Xue et al. [83] describes microtubules, neurofilaments, and organelles as a system of stochastically interacting particles in an axonal cross-section. It builds upon particle binding, unbinding, addition, and removal that occur on a time scale of seconds or fractions of a second and addresses polymer segregation that occurs on a time scale of hours. Simulations of the model suggest that microtubule zippering by organelles is the main mechanism for the cytoskeleton segregation. The zipping mechanism was further investigated using a more detailed model in Sect. 3.2.

The model in Xue et al. [83] brought up several open problems for mathematical modeling and analysis. Due to the vast span of time scales, the model in [83] has to be simulated on a time scale of milliseconds to ensure accuracy and thus is computationally expensive when the particle number becomes large. To overcome this difficulty, coarse-grained models need to be derived by averaging out the fast events. This requires the development of new asymptotic analysis for stochastically interacting particle systems. To achieve this goal, one could start with a simplified stochastic model which replaces the indirect interaction of microtubules through the motor-organelle complexes by direct, long-range stochastic forces between microtubule pairs. A different route is to develop fast stochastic simulation methods for stochastically interacting particle systems that have several intrinsic time scales.

It has been observed that if treated by IDPN, the axonal cytoskeleton separates along the whole length of the axon, and if IDPN is washed out before visible axonal swelling, microtubule–neurofilament remixing occurs as a wave that propagates from the cell body to the axon terminal at a speed of 1–2 mm per day [61]. Microtubules are long polymers that can span tens of microns in axons. In the

absence of neurofilament transport, how do organelles zip microtubules together as they move along multiple microtubules? With regard to the remixing stage, is the proximo-distal remixing a result of the transport of newly synthesized neurofilaments from the cell body? What determines the wave speed? How does axonal swelling occur? One approach to address these questions is to construct a 3D stochastic model that extends [83] and Sect. 3.2. The model could describe microtubules as semi-flexible polymers, each of which modeled as a bead-spring system, and neurofilaments and organelles as cargoes with a fixed volume that bind, unbind, and move along microtubules randomly. This approach is computationally expensive but allows us to tightly integrate the biophysical properties of the cytoskeleton. Simplified stochastic models that are computationally more affordable can also be constructed to gain insights into the problem.

PDE models in 3D can be constructed to describe the distribution and movement of microtubules, neurofilaments, and organelles inside the cytosol. These PDEs can be derived either using heuristic arguments or, if possible, from the aforementioned stochastic models. This is a challenging task because the underlying problem involves fluid–structure interaction and active bidirectional cargo transport, in addition to the particle volume exclusion. The PDE models can be used to address a variety of phenomena related to axonal transport both in healthy axons and in diseased axons.

5 Summary

New experimental methods have advanced our understanding of the molecular mechanisms of intracellular transport in vitro, and new imaging techniques have allowed us to observe single cargo movement in living cells. This has significantly advanced our understanding regarding how various cargoes are transported to their destinations inside an axon. The integration of experimental data is required to address questions at the system level, e.g., how intracellular traffic is regulated inside the crowded axon, how axonal transport affects the development and maintenance of the shape of an axon, and how phenomena observed at different space and time scales are related. To achieve this goal, quantitative models and multiscale methods for axonal transport and traffic problems must be developed.

In this chapter, we have summarized recent progress on mathematical modeling of axonal transport. In Sect. 2, we reviewed mathematical models and analyses that describe the distribution of a single type of cargo, either organelles or macromolecular proteins, along the length of a normal axon. These models were motivated by 1D experimental data on the population density of cargoes collected using pulse-labeling method. In Sect. 3, we reviewed a recent stochastic model that advanced our understanding of the underlying mechanisms of axonal swelling and cytoskeleton reorganization observed in neurological disorders. We further

investigated the causal mechanism of the segregation phenomena, microtubule zippering by moving organelles, using a more detail model. In Sect. 4, we discussed several open problems in this exciting area, which require advances in stochastic and PDE methods.

Acknowledgements This research was supported by US NSF DMS 1312966 and US NSF CAREER Award 1553637. CX was also supported by the Mathematical Biosciences Institute as a long-term visitor.

References

1. M.G.L. Van den Heuvel, M.P. de Graaff, C. Dekker, Microtubule curvatures under perpendicular electric forces reveal a low persistence length. Proc. Natl. Acad. Sci. U. S. A. **105**(23), 7941–7946 (2008)
2. S. Tsukita, H. Ishikawa, The cytoskeleton in myelinated axons: serial section study. Biomed. Res. **2**, 424–437 (1981)
3. B.J. Schnapp, T.S. Reese, Cytoplasmic structure in rapid frozen axons. J. Cell Biol. **94**, 667–679 (1982)
4. N. Hirokawa, Axonal transport and the cytoskeleton. Curr. Opin. Neurobiol. **3**(5), 724–731 (1993)
5. A. Brown, Axonal transport, in *Neuroscience in the 21st Century* (Springer, Berlin, 2013)
6. N. Hirokawa, R. Takemura, Molecular motors and mechanisms of directional transport in neurons. Nat. Rev. Neurosci. **6**(3), 201–214 (2005)
7. F. Gittes, B. Mickey, J. Nettleton, J. Howard, Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. J. Cell Biol. **120**(4), 923–934 (1993)
8. H. Isambert, P. Venier, A.C. Maggs, A. Fattoum, R. Kassab, D. Pantaloni, M.F. Carlier, Flexibility of actin filaments derived from thermal fluctuations. effect of bound nucleotide, phalloidin, and muscle regulatory proteins. J. Biol. Chem. **270**(19), 11437–11444 (1995)
9. K. Xu, G. Zhong, X. Zhuang, Actin, spectrin, and associated proteins form a periodic cytoskeletal structure in axons. Science **339**(6118), 452–456 (2013)
10. E.L. Bearer, T.S. Reese, Association of actin filaments with axonal microtubule tracts. J. Neurocytol. **28**(2), 85–98 (1999)
11. P.J. Hollenbeck, W.M. Saxton, The axonal transport of mitochondria. J. Cell Sci. **118**(Pt 23), 5411–5419 (2005)
12. R. Beck, J. Deek, M. C. Choi, T. Ikawa, O. Watanabe, E. Frey, P. Pincus, C.R. Safinya, Unconventional salt trend from soft to stiff in single neurofilament biopolymers. Langmuir **26**(24), 18595–18599 (2010)
13. O.I. Wagner, S. Rammensee, N. Korde, Q. Wen, J.-F. Leterrier, P.A. Janmey, Softness, strength and self-repair in intermediate filament networks. Exp. Cell Res. **313**(10), 2228–2235 (2007)
14. R. Perrot, P. Lonchampt, A.C. Peterson, J. Eyer, Axonal neurofilaments control multiple fiber properties but do not influence structure or spacing of nodes of Ranvier. J. Neurosci. **27**(36), 9573–9584 (2007)
15. R.L. Friede, T. Samorajski, Axon caliber related to neurofilaments and microtubules in sciatic nerve fibers of rats and mice. Anat. Rec. **167**(4), 379–387 (1970)
16. L. Wang, C.L. Ho, D. Sun, R.K. Liem, A. Brown, Rapid movement of axonal neurofilaments interrupted by prolonged pauses. Nat. Cell Biol. **2**(3), 137–141 (2000)
17. N. Trivedi, P. Jung, A. Brown, Neurofilaments switch between distinct mobile and stationary states during their transport along axons. J. Neurosci. **27**(3), 507–516 (2007)
18. E. Chevalier-Larsen, E.L.F. Holzbaur, Axonal transport and neurodegenerative disease. Biochim. Biophys. Acta **1762**(11–12), 1094–1108 (2006)

19. K.J. De Vos, A.J. Grierson, S. Ackerley, C.C.J. Miller, Role of axonal transport in neurodegenerative diseases. Annu. Rev. Neurosci. **31**, 151–173 (2008)
20. S. Millecamps, J.-P. Julien, Axonal transport deficits and neurodegenerative diseases. Nat. Rev. Neurosci. **14**(3), 161–176 (2013)
21. C. Zhao, J. Takita, Y. Tanaka, M. Setou, T. Nakagawa, S. Takeda, H.W. Yang, S. Terada, T. Nakata, Y. Takei, M. Saito, S. Tsuji, Y. Hayashi, N. Hirokawa, Charcot-Marie-tooth disease type 2A caused by mutation in a microtubule motor kIF1Bbeta. Cell **105**(5), 587–597 (2001)
22. L. Wang, A. Brown, A hereditary spastic paraplegia mutation in kinesin-1A/KIF5A disrupts neurofilament transport. Mol. Neurodegener. **5**, 52 (2010)
23. M.E. MacDonald, C.M. Ambrose, M.P. Duyao, R.H. Myers, C. Lin, L. Srinidhi, G. Barnes, S.A. Taylor, M. James, N. Groot et al., A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell **72**(6), 971–983 (1993)
24. M. Katsuno, H. Adachi, M. Minamiyama, M. Waza, K. Tokui, H. Banno, K.e Suzuki, Y. Onoda, F. Tanaka, M. Doyu, G. Sobue, Reversible disruption of dynactin 1-mediated retrograde axonal transport in polyglutamine-induced motor neuron degeneration. J. Neurosci. **26**(47), 12106–12117 (2006)
25. I. Puls, C. Jonnakuty, B.H. LaMonte, E.L.F. Holzbaur, M. Tokito, E. Mann, M.K. Floeter, K. Bidus, D. Drayna, S.J. Oh, R.H. Brown Jr., C.L. Ludlow, K.H. Fischbeck, Mutant dynactin in motor neuron disease. Nat. Genet. **33**(4), 455–456 (2003)
26. G.M. Fabrizi, T. Cavallaro, C. Angiari, I. Cabrini, F. Taioli, G. Malerba, L. Bertolasi, N. Rizzuto, Charcot-Marie-tooth disease type 2E, a disorder of the cytoskeleton. Brain **130**(Pt 2), 394–403 (2007)
27. D.D. Tshala-Katumbay, V.S. Palmer, M.R. Lasarev, R.J. Kayton, M.I. Sabri, P.S. Spencer, Monocyclic and dicyclic hydrocarbons: structural requirements for proximal giant axonopathy. Acta Neuropathol. **112**(3), 317–324 (2006)
28. H.H. Goebel, P. Vogel, M. Gabriel, Neuropathologic and morphometric studies in hereditary motor and sensory neuropathy type II with neurofilament accumulation. Ital J. Neurol. Sci. **7**(3), 325–332 (1986)
29. A.L. Taratuto, G. Sevlever, M. Saccoliti, L. Caceres, M. Schultz, Giant axonal neuropathy (GAN): an immunohistochemical and ultrastructural study report of a Latin American case. Acta Neuropathol. **80**(6), 680–683 (1990)
30. M. Donaghy, R.H. King, P.K. Thomas, J.M. Workman, Abnormalities of the axonal cytoskeleton in giant axonal neuropathy. J. Neurocytol. **17**(2), 197–208 (1988)
31. I.R. Griffiths, I.D. Duncan, M. McCulloch, S. Carmichael, Further studies of the central nervous system in canine giant axonal neuropathy. Neuropathol. Appl. Neurobiol. **6**(6), 421–432 (1980)
32. R. S. Smith. The short term accumulation of axonally transported organelles in the region of localized lesions of single myelinated axons. J. Neurocytol. **9**(1), 39–65 (1980)
33. P.C. Bressloff, J.M. Newby, Stochastic models of intracellular transport. Rev. Mod. Phys. **85**(1), 135 (2013)
34. D. Chowdhury, Stochastic mechano-chemical kinetics of molecular motors: a multidisciplinary enterprise from a physicist's perspective. Phys. Rep. **529**(1), 1–197 (2013)
35. A. Brown, Slow axonal transport. New Encycl. Neurosci. **9**, 1–9 (2009)
36. S.I. Rubinow, J.J. Blum, A theoretical approach to the analysis of axonal transport. Biophys. J. **30**(1), 137–147 (1980)
37. T. Takenaka, H. Gotoh, Simulation of axoplasmic transport. J. Theor. Biol. **107**(4), 579–601 (1984)
38. J.J. Blum, M.C. Reed, A model for fast axonal transport. Cell Motil. **5**(6), 507–527 (1985)
39. M.C. Reed, J.J. Blum, Theoretical analysis of radioactivity profiles during fast axonal transport: effects of deposition and turnover. Cell Motil. Cytoskeleton **6**(6), 620–627 (1986)
40. J.J. Blum, M.C. Reed, A model for slow axonal transport and its application to neurofilamentous neuropathies. Cell Motil. Cytoskeleton **12**(1), 53–65 (1989)
41. M.C. Reed, S. Venakides, J.J. Blum, Approximate traveling waves in linear reaction-hyperbolic equations. SIAM J. Appl. Math. **50**(1), 167–180 (1990)

42. S. Roy, P. Coffee, G. Smith, R.K. Liem, S.T. Brady, M.M. Black, Neurofilaments are transported rapidly but intermittently in axons: implications for slow axonal transport. J. Neurosci. **20**(18), 6849–6861 (2000)
43. G. Craciun, A. Brown, A. Friedman, A dynamical system model of neurofilament transport in axons. J. Theor. Biol. **237**(3), 316–322 (2005)
44. A. Brown, L. Wang, P. Jung, Stochastic simulation of neurofilament transport in axons: the "stop-and-go" hypothesis. Mol. Biol. Cell **16**(9), 4243–4255 (2005)
45. P. Jung, A. Brown, Modeling the slowing of neurofilament transport along the mouse sciatic nerve. Phys. Biol. **6**(4), 046002 (2009)
46. Y. Li, P. Jung, A. Brown, Axonal transport of neurofilaments: a single population of intermittently moving polymers. J. Neurosci. **32**(2), 746–758 (2012)
47. Y. Li, A. Brown, P. Jung, Deciphering the axonal transport kinetics of neurofilaments using the fluorescence photoactivation pulse-escape method. Phys. Biol. **11**(2), 026001 (2014)
48. P.C. Monsma, Y. Li, J.D. Fenn, P. Jung, A. Brown, Local regulation of neurofilament transport by myelinating cells. J. Neurosci. **34**(8), 2979–2988 (2014)
49. A. Friedman, G. Craciun, Approximate traveling waves in linear reaction-hyperbolic equations. SIAM J. Math. Anal. **38**(3), 741–758 (2006)
50. A. Friedman, G. Craciun, A model of intracellular transport of particles in an axon. J. Math. Biol. **51**(2), 217–246 (2005)
51. A. Friedman, B. Hu, Uniform convergence for approximate traveling waves in linear reaction–diffusion–hyperbolic systems. Arch. Ration. Mech. Anal. **186**(2), 251–274 (2007)
52. A. Friedman, B. Hu, Uniform convergence for approximate traveling waves in linear reaction-hyperbolic systems. Indiana University Math. J. **56**(5), 2133–2158 (2007)
53. E.A. Brooks, Probabilistic methods for a linear reaction-hyperbolic system with constant coefficients. Ann. Appl. Probab. **9**(3), 719–731 (1999)
54. L. Popovic, S.A. McKinley, M.C. Reed, A stochastic compartmental model for fast axonal transport. SIAM J. Appl. Math. **71**(4), 1531–1556 (2011)
55. P.C. Bressloff, Stochastic model of protein receptor trafficking prior to synaptogenesis. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **74**(3 Pt 1), 031910 (2006)
56. J.M. Newby, P.C. Bressloff, Directed intermittent search for a hidden target on a dendritic tree. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **80**(2 Pt 1), 021913 (2009)
57. J.M. Newby, P.C. Bressloff, Quasi-steady state reduction of molecular motor-based models of directed intermittent search. Bull. Math. Biol. **72**(7), 1840–1866 (2010)
58. P.C. Bressloff, J.M. Newby, Stochastic hybrid model of spontaneous dendritic NMDA spikes. Phys. Biol. **11**(1), 016006 (2014)
59. R.L. Price, P. Paggi, R.J. Lasek, M.J. Katz, Neurofilaments are spaced randomly in the radial dimension of axons. J. Neurocytol. **17**(1), 55–62 (1988)
60. S.T. Hsieh, G.J. Kidd, T.O. Crawford, Z. Xu, W.M. Lin, B.D. Trapp, D.W. Cleveland, J.W. Griffin, Regional modulation of neurofilament organization by myelination in normal axons. J. Neurosci. **14**(11 Pt 1), 6392–6401 (1994)
61. S.C. Papasozomenos, L. Autilio-Gambetti, P. Gambetti, Reorganization of axoplasmic organelles following beta, beta′-iminodipropionitrile administration. J. Cell Biol. **91**(3 Pt 1), 866–871 (1981)
62. S.C. Papasozomenos, M. Yoon, R. Crane, L. Autilio-Gambetti, P. Gambetti, Redistribution of proteins of fast axonal transport following administration of beta,beta′-iminodipropionitrile: a quantitative autoradiographic study. J. Cell Biol. **95**(2 Pt 1), 672–675 (1982)
63. J.W. Griffin, K.E. Fahnestock, D.L. Price, P.N. Hoffman, Microtubule-neurofilament segregation produced by beta, beta′-iminodipropionitrile: evidence for the association of fast axonal transport with microtubules. J. Neurosci. **3**(3), 557–566 (1983)
64. J.W. Griffin, K.E. Fahnestock, D.L. Price, L.C. Cork, Cytoskeletal disorganization induced by local application of β,β'-iminodipropionitrile and 2,5-hexanedione. Ann. Neurol. **14**, 55–61 (1983)
65. J.W. Griffin, D.L. Price, P.N. Hoffman, Neurotoxic probes of the axonal cytoskeleton. Trends Neurosci. **6**, 490–495 (1983)

66. S.C. Papasozomenos, L.I. Binder, P.K. Bender, M.R. Payne, Microtubule-associated protein 2 within axons of spinal motor neurons: associations with microtubules and neurofilaments in normal and beta,beta′-iminodipropionitrile-treated axons. J. Cell Biol. **100**(1), 74–85 (1985)
67. N. Hirokawa, G.S. Bloom, R.B. Vallee, Cytoskeletal architecture and immunocytochemical localization of microtubule-associated proteins in regions of axons associated with rapid axonal transport: the beta,beta′-iminodipropionitrile-intoxicated axon as a model system. J. Cell Biol. **101**(1), 227–239 (1985)
68. S.C. Papasozomenos, M.R. Payne, Actin immunoreactivity localizes with segregated microtubules and membranous organelles and in the subaxolemmal region in the beta,beta′-iminodipropionitrile axon. J. Neurosci. **6**(12), 3483–3491 (1986)
69. R.G. Nagele, K.T. Bush, H.Y. Lee, A morphometric study of cytoskeletal reorganization in rat sciatic nerve axons following beta,beta′-iminodipropionitrile (IDPN) treatment. Neurosci. Lett. **92**(3), 241–246 (1988)
70. A. Bizzi, R.C. Crane, L. Autilio-Gambetti, P. Gambetti, Aluminum effect on slow axonal transport: a novel impairment of neurofilament transport. J. Neurosci. **4**(3), 722–731 (1984)
71. M.R. Gottfried, D.G. Graham, M. Morgan, H.W. Casey, J.S. Bus, The morphology of carbon disulfide neurotoxicity. Neurotoxicology **6**(4), 89–96 (1985)
72. I. Jirmanová, E. Lukás, Ultrastructure of carbon disulphie neuropathy. Acta Neuropathol. **63**(3), 255–263 (1984)
73. Z. Sahenk, J.R. Mendell, Alterations in slow transport kinetics induced by estramustine phosphate, an agent binding to microtubule-associated proteins. J. Neurosci. Res. **32**(4), 481–493 (1992)
74. D.D. Tshala-Katumbay, V.S. Palmer, R.J. Kayton, M.I. Sabri, P.S. Spencer, A new murine model of giant proximal axonopathy. Acta Neuropathol. **109**(4), 405–410 (2005)
75. M.K. Lee, J.R. Marszalek, D.W. Cleveland, A mutant neurofilament subunit causes massive, selective motor neuron death: implications for the pathogenesis of human motor neuron disease. Neuron **13**(4), 975–988 (1994)
76. H. Selye, Lathyrism. Rev. Can. Biol. **16**, 1–73 (1957)
77. J.L. Cadet, The iminodipropionitrile (IDPN)-induced dyskinetic syndrome: behavioral and biochemical pharmacology. Neurosci. Biobehav. Rev. **13**(1), 39–45 (1989)
78. P.S. Spencer, H.H. Schaumburg, Lathyrism: a neurotoxic disease. Neurobehav. Toxicol. Teratol. **5**(6), 625–629 (1983)
79. P.S. Spencer, C.N. Allen, G.E. Kisby, A.C. Ludolph, S.M. Ross, D.N. Roy, Lathyrism and western pacific amyotrophic lateral sclerosis: etiology of short and long latency motor system disorders. Adv. Neurol. **56**, 287–299 (1991)
80. J. Llorens, C. Soler-Martín, B. Cutillas, S. Saldaña-Ruíz, Nervous and vestibular toxicities of acrylonitrile and iminodipropionitrile. Toxicol. Sci. **110**(1), 244–245; Author reply 246–248 (2009)
81. J. Llorens, Toxic neurofilamentous axonopathies – accumulation of neurofilaments and axonal degeneration. J. Intern. Med. **273**(5), 478–489 (2013)
82. J.W. Griffin, P.N. Hoffman, A.W. Clark, P.T. Carroll, D.L. Price, Slow axonal transport of neurofilament proteins: impairment of beta,beta′-iminodipropionitrile administration. Science **202**(4368), 633–635 (1978)
83. C. Xue, B. Shtylla, A. Brown, A stochastic multiscale model that explains the segregation of axonal microtubules and neurofilaments in toxic neuropathies. PLoS Comput. Biol. **11**(8), e1004406 (2015)
84. Q. Zhu, M. Lindenbaum, F. Levavasseur, H. Jacomy, J.P. Julien, Disruption of the NF-H gene increases axonal microtubule content and velocity of neurofilament transport: relief of axonopathy resulting from the toxin beta,beta′-iminodipropionitrile. J. Cell Biol. **143**(1), 183–193 (1998)
85. K. Visscher, M.J. Schnitzer, S.M. Block, Single kinesin molecules studied with a molecular force clamp. Nature **400**(6740), 184–189 (1999)
86. R.P. Erickson, Z. Jia, S.P. Gross, C.C. Yu, How molecular motors are arranged on a cargo is important for vesicular transport. PLoS Comput. Biol. **7**(5), e1002032 (2011)

87. N. Hirokawa, K.K. Pfister, H. Yorifuji, M.C. Wagner, S.T. Brady, G.S. Bloom, Submolecular domains of bovine brain kinesin identified by electron microscopy and monoclonal antibody decoration. Cell **56**(5), 867–878 (1989)
88. C.M. Coppin, J.T. Finer, J.A. Spudich, R.D. Vale, Measurement of the isometric force exerted by a single kinesin molecule. Biophys. J. **68**(4 Suppl.), 242S–244S (1995)
89. F. Ziebert, M. Vershinin, S.P. Gross, I.S. Aranson, Collective alignment of polar filaments by molecular motors. Eur. Phys. J. E Soft Matter **28**(4), 401–409 (2009)
90. R.G. Cox, The motion of long slender bodies in a viscous fluid. Part 1. general theory. J. Fluid Mech. **44**(Part 3), 790–810 (1970)
91. C. Brennen, H. Winet, Fluid mechanics of propulsion by cilia and flagella. Annu. Rev. Fluid Mech. **9**(1), 339–398 (1977)
92. K. Svoboda, C.F. Schmidt, B.J. Schnapp, S.M. Block, Direct observation of kinesin stepping by optical trapping interferometry. Nature **365**(6448), 721–727 (1993)
93. M.J. Schnitzer, S.M. Block, Kinesin hydrolyses one ATP per 8-nm step. Nature **388**(6640), 386–390 (1997)
94. A. Kunwar, M. Vershinin, J. Xu, S.P. Gross, Stepping, strain gating, and an unexpected force-velocity curve for multiple-motor-based transport. Curr. Biol. **18**(16), 1173–1183 (2008)
95. J. Kerssemakers, J. Howard, H. Hess, S. Diez, The distance that kinesin-1 holds its cargo from the microtubule surface measured by fluorescence interference contrast microscopy. Proc. Natl. Acad. Sci. U. S. A. **103**(43), 15812–15817 (2006)
96. I.G Currie, *Fundamental Mechanics of Fluids* (CRC Press, Boca Raton, 2012)
97. R. Swaminathan, C.P. Hoang, A.S. Verkman, Photobleaching recovery and anisotropy decay of green fluorescent protein GFP-s65t in solution and cells: cytoplasmic viscosity probed by green fluorescent protein translational and rotational diffusion. Biophys. J. **72**, 1900–1907 (1997)

Stochastic Models for Evolving Cellular Populations of Mitochondria: Disease, Development, and Ageing

Hanne Hoitzing, Iain G. Johnston, and Nick S. Jones

1 Introduction

About 2 billion years ago, a bacterium was engulfed by another cell as an endosymbiont [1]. The relationship between bacterium and cell turned out to be a beneficial one, perhaps as the bacterium provided a new source of energy for the cell, which in turn provided protection from the environment. Over millions of years of evolution, the bacterium lost its independence and became an organelle of the cell: the mitochondrion. Currently most cells in our body cannot survive without mitochondria which, besides being the main energy producers in the cell, are involved in various other processes including intracellular calcium signalling, iron–sulphur cluster biogenesis, and cell death [2–5]. Mitochondria can be highly dynamic organelles, continuously undergoing fusion and fission events which leads to a diverse range of mitochondrial morphologies, from fragmented states to continuous networks [6, 7]. Correctly balancing mitochondrial fusion and fission is important for cellular functionality [8, 9], and dysfunctions in mitochondrial fusion–fission dynamics have been observed in numerous diseases [7, 10–14]. Models of mitochondrial fusion–fission and implications on cellular health are discussed in Sect. 3.2.

Due to their rich evolutionary history [15], mitochondria retain their own genetic material: mitochondrial DNA (mtDNA). MtDNA is tiny compared to nuclear DNA and in humans comprises only 16,569 base pairs, encoding only 37 genes. The number of mtDNA molecules in a cell depends on the type of cell, and can vary over

H. Hoitzing • N.S. Jones (✉)
Imperial College London, South Kensington Campus, London SW7 2AZ, UK
e-mail: nick.jones@imperial.ac.uk

I.G. Johnston
School of Biosciences, University of Birmingham, Birmingham B15 2TT, UK

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_13

time. Replication of mtDNA can occur independent of the cell cycle [16] (though it is linked to certain stages of the cell cycle, see, e.g., [17]), ensuring continuous turnover of the mtDNA population in most dividing as well as non-dividing cells. Errors can occur during replication of mtDNA molecules. Mitochondria do contain machinery to repair mtDNA [18, 19], but this machinery is possibly less efficient than nuclear DNA repair. When mutations occur, they often coexist with wildtype mtDNA molecules, a situation which is called *heteroplasmy*. Denoting the number of wildtype and mutant mtDNA molecules in a cell by w and m, respectively, the level of heteroplasmy is defined as

$$h = \frac{m}{w+m} \tag{1}$$

and can vary between cells in the same tissue, between different tissues, and between individuals.

Mutations in mtDNA can have detrimental consequences, and mitochondrial diseases (including mutations of both mtDNA and nuclear DNA) affect $\sim$1 in 4300 of the adult human population [20]. Because of the large number of mtDNAs in a cell, the presence of a few mutants does not immediately cause major problems. The heteroplasmy value has to exceed a critical threshold, typically 60–90%, before biochemical defects are observed [21–24]. Interestingly, the same pathogenic mutations that cause mtDNA diseases are also found in healthy individuals but at much lower levels [25, 26]. Can these low frequency mutants suddenly expand? Are there ways in which we can prevent them from doing so? Understanding the dynamics of mtDNA molecules inside cells and the way in which mutants can accumulate over time, or even take over the whole mtDNA population (homoplasmy), is a crucial step in understanding the progression of mtDNA diseases [66, 115, 120] and diseases which might be linked to mitochondria (e.g. cancer [11], Parkinson's [10], diabetes [7], and Alzheimer's [12]).

Because the severity of mitochondrial diseases is partly related to the proportion of cells that have heteroplasmy values above the critical threshold, it is important to have knowledge of how the heteroplasmy *distribution* changes over time. Stochastic modelling allows us to investigate these important distributions and analyse the cell-to-cell variability in heteroplasmy, as opposed to deterministic models which typically only describe mean behaviour.

Mutations (pathological and non-pathological) accumulate with age in any healthy individual. Mutations associated with ageing are typically seen in post-mitotic tissues (e.g. myocardium and brain) and often a cell contains very high fractions of a single mutation, which is said to have clonally expanded [27]. Different cells usually contain distinct mutations though some types of mutations occur more often (like the 'common deletion', e.g. [28]). It is not yet clear how clonal expansion arises, and there exist several hypotheses on this topic.

Many of the hypotheses on mutant clonal expansion involve a selection advantage for the mutant species such as: a shorter replication time for deletion mutants because of their smaller genomes [29–31]; the survival-of-the-slowest (SOS) hypothesis which assumes that certain mutations reduce the release of

damaging superoxide molecules, with the result that these mutant mitochondria are less often degraded than wildtype mitochondria [32, 33]; the 'crippled mitochondrion' hypothesis which states that mitochondrial biogenesis is partly controlled by the mitochondrion itself, and that mutant mtDNA molecules create a microenvironment that stimulates their own replication [34, 35]. Even though all of these hypotheses have some attractive features, none of them are fully supported by experimental data (a critical review is given in [42]). A recently proposed selection mechanism involves a faster replication rate for mutants caused by differences in transcription rates [36] and is discussed in more detail in Sect. 4.

One of the hypotheses that does not involve a direct selection advantage for mutants is the vicious cycle hypothesis. Unlike the SOS hypothesis, it states that mutants create *more* damaging radicals and thereby create more mutants which then create more radicals, forming a vicious cycle [37–39]. The vicious cycle hypothesis predicts a whole range of different mutations to occur, which is not seen experimentally and evidence points towards replication as the major source of errors [40]. Another hypothesis is that stochastic drift of mtDNA molecules over time can account for the observed clonal expansion. In theory, mutants can take over entire mtDNA populations purely by chance because of the stochasticity of mtDNA replication and degradation, and cell divisions. This idea is also not fully supported by data, as versions of it cannot explain clonal expansion in short-lived animals [41]. Currently a debate exists in the literature as to whether selective differences exist between types of mtDNA in mixed populations [41–44].

Studying the dynamics of mtDNA molecules and mutant accumulation (with or without additional selection advantages for mutants), and the effects of cell divisions and possible nuclear feedback mechanisms is an important step in understanding the ageing process and the progression of mtDNA diseases. There have been numerous models describing mtDNA dynamics in individual non-dividing cells, in dividing cells, on a tissue level, or across generations (reviewed below). Both the physical (fusion–fission) and the genetic (mtDNA dynamics) properties of mitochondria are linked and stochastic, meaning that physical and genetic stochastic models are valuable. We will discuss the different type of models that are considered and some of their main results.

2 Experimental Observations

In healthy cells, mtDNA levels are controlled and remain fairly constant over time as we age [45]. The number of mtDNA molecules per cell ranges from about 100 to 10^5 and depends heavily on the type of cell (e.g. mtDNA levels per cell have been measured to be 3650 ± 620 in skeletal muscle, 6970 ± 920 in heart [45], and mature human oocytes have around 10^5 mtDNA copies). In heteroplasmic cells, i.e. cells with both wildtype and mutant mtDNA, total mtDNA copy number can be 5- to 17-fold higher than in cells with only wildtype mtDNA [35, 46–48], and this proliferation is one of the hallmarks of certain heteroplasmic mtDNA mutations.

As we age, heteroplasmy levels can start to vary dramatically between tissues [49, 50], being particularly high in the putamen, cerebral cortex and substantia nigra [51, 52]. Mammalian aged tissues show a mosaic pattern of healthy cells and severely dysfunctional cells, meaning that a patch of healthy cells can occur directly adjacent to cells with high mutant loads (reviewed in [53]). A small number of very dysfunctional cells can sometimes have large effects on tissue performance, and this non-linearity provides another reason why studying distributions rather than mean behaviours is important. By the eighth decade of life, $\leq$5% of postmitotic cells develop COX deficiency [54–56]. By the eighth decade of life, $\sim(0.1-5)\%$ of postmitotic cells develop mitochondrial deficiencies due to high mutant levels [54–56]. Perhaps surprisingly, rodents show similar levels of deficiency at only 3 years [57, 58].

Experimental measurements of parameters that are often used in stochastic models of mtDNA dynamics are summarized in Table 1. Note that interpreting experimental data can be challenging because sources of uncertainty are introduced through experimental measurement [59]. A Bayesian model was constructed to partially address this problem [60], inferring posterior parameter distributions for the substantia nigra region of the human brain (Table 1).

Table 1 Measured values (or values often used in models) of parameters relevant for mathematical models of mitochondria

Quantity of interest	Measured value(s)
Critical threshold that mtDNA mutations have to pass in order for the cell to show biochemical defects	– >60% for single large mtDNA deletions [61] – >90% for certain point mutations in tRNA genes [62] – a mean of 96.2% with a 95% confidence interval of (86.8%, 99.9%) for mtDNA deletions in substantia nigra neurons [60]
Probability of a mutation event during mtDNA replication	– $10^{-3} - 10^{-6}$, used in model [43] – likely to be between 10^{-4} and 10^{-5} based on data from substantia nigra neurons [60]
Half-life of mtDNA molecules	On the order of 1–3 weeks [63–65], but highly tissue dependent [66]
The percentage of dysfunctional cells in a tissue with high mutant loads at the end of an organism's lifespan	– A type of focal respiratory chain deficiency is found in >15% of all colonic crypts of humans older than 80 years [67] – About 10% is often used in models [36, 68]
Time required for mtDNA replication	– 1–2 h [69, 70]

3 Stochastic Modelling of Mitochondria

Our coverage of stochastic models of mitochondria starts by introducing a well-known and often-used model of mtDNA dynamics, the 'relaxed replication model'. This model is discussed in some detail because it gives an intuitive possible explanation for the observed threshold effect and it is referred to by many subsequent models. Afterwards, a brief discussion of other *in silico* models of mtDNA dynamics and mitochondrial fusion–fission dynamics is given. Finally, an analytical model of mtDNA dynamics is discussed, which generalizes the relaxed replication model by allowing for arbitrary nuclear control of the replication and degradation rates of mtDNA. The specific role of mtDNA dynamics in ageing and development is discussed in the next section. We note briefly the types of stochastic mitochondrial phenomena that we do not consider: variability in mitochondrial network structure independent of genetics, e.g. [8, 71, 72], organ-to-organ variability, e.g. [66], how variability in mitochondrial content and function might link to gene expression variability, e.g. [73], and how individual membrane potentials might fluctuate, e.g. [74].

3.1 Relaxed Replication and Its Implications

In 1999, a stochastic model was developed to study how populations of mtDNA molecules vary over time [75]. This model of mtDNA dynamics is known as 'relaxed replication' because it assumes mtDNA turnover occurs continuously over time, independent of cell division. It has been subsequently used in a variety of other models, e.g. [41, 43, 76] and has obtained experimental support, e.g. [77]. Two situations were investigated, both of which concerned simulations of cells with two different types of mtDNA molecules. In the first case both types were neutral and in the second case one of the types was pathological. The aim was to see how the presence of mutant molecules affects the overall dynamics, and how the fraction of mutants varies over time. The population of mtDNA molecules is assumed to be well-mixed and cells are assumed to be non-dividing.

In the case of two neutral types of mtDNA molecules, the cell tries to meet a certain copy number, and the main dynamics are described by the following ODE:

$$\begin{aligned} \frac{dN}{dt} &= C - \mu N \\ &= \mu(N_{\text{opt}} - N) \end{aligned} \tag{2}$$

where N is the total number of mtDNA molecules, C is the copy rate at which new mtDNA molecules are generated, and μ is their degradation rate. The constant C is chosen such that the total population is controlled towards a desired value N_{opt}. When embedded in a stochastic framework, the above equation describes an

Immigration-Death model with constant immigration (C) and death (μ) rates. The corresponding master equation for two neutral alleles A and B (with $N = A + B$) is given by:

$$\frac{\partial P_{AB}(t)}{\partial t} = \mu N_{\text{opt}}\Big(P_{A-1,B}(t) + P_{A,B-1}(t)\Big) + \mu(A+B+1)\Big(P_{A+1,B}(t) + P_{A,B+1}(t)\Big) - 2\mu(N_{\text{opt}} + A + B)P_{AB}(t) \tag{3}$$

Master equation descriptions of mtDNA populations can sometimes be solved explicitly [102, 107], but this is often not the case and approximation methods must be made, such as the system size expansion (discussed in Sect. 3.3). The original implementation of this model combines deterministic immigration events with stochastic death events, which is arguably less natural than a full stochastic model (see supplement of [78]). The initial conditions were taken such that $A(0) + B(0) = N_{\text{opt}}$, i.e. the system started in steady state. Simulations showed that, on average, the proportions of alleles A and B remain constant. The probability that a certain allele takes over the entire population was found to be equal to the initial allele frequency, consonant with a random drift model. The relaxed replication model is often referred to as 'the random drift model' (though the dynamics are non-trivial), and this term has now come to refer to models with an absence of selective differences between mutant and wildtype species.

To include the presence of mutant mtDNA, the dynamics in Eq. (2) were slightly changed to $dN/dt = C(N) - \mu N$, i.e. the replication rate now depends on the state of the system. It was argued that a severely pathological mutant should not contribute to the replication feedback, meaning that $C(N) = C(w)$, i.e. the control is only dependent on the wildtype species. The replication rate was then multiplied by the fraction of the species (proportional selection) which leads to:

$$\begin{aligned} \frac{dw}{dt} &= C(w)\frac{w}{w+m} - \mu w \\ \frac{dm}{dt} &= C(w)\frac{m}{w+m} - \mu m \\ C(w) &= \mu N_{\text{opt}}\left(\alpha + (1-\alpha)\frac{w}{N_{\text{opt}}}\right) \end{aligned} \tag{4}$$

The parameter $\alpha \geq 1$ describes the response of the system. The idea is that the cell still tries to maintain the same number of wildtype mtDNAs as it would when no mutants are present, i.e. when $w = N_{\text{opt}}$, $C(w) = \mu N_{\text{opt}}$ as it was in Eq. (2).

Using Eq. (4), the deterministic steady states of the system [denoted by (w_s, m_s)] can easily be found, and they form a line defined by

$$w_s + \frac{1}{\alpha} m_s = N_{\text{opt}} \tag{5}$$

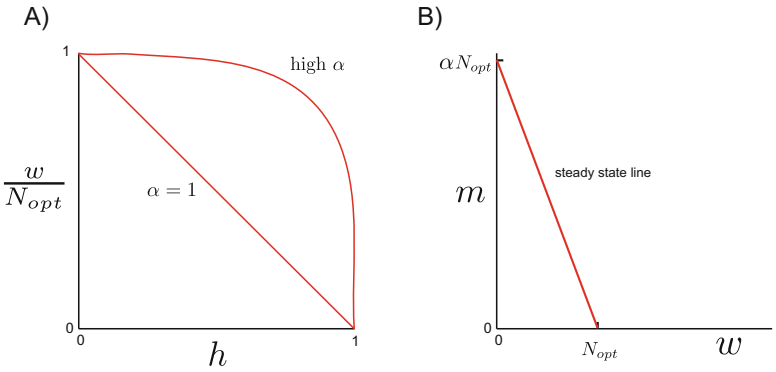


Fig. 1 Steady state lines of the relaxed replication model. (**a**) The number of wildtype mtDNA (relative to the desired value N_{opt}) as a function of α. For high α, w remains close to N_{opt} for a wide range of heteroplasmies until it suddenly drops at high h, creating an effective threshold effect. The functions that are plotted are described by $w_s/N_{\text{opt}} = (1-h)/(1-(1-1/\alpha)h)$. This behaviour has been observed in skeletal muscle fibres [77]. (**b**) The line of steady states, showing that a large value for α results in a large number of mutants if h is high

Stochastic events will cause trajectories to fluctuate around the deterministic steady state line until one of the absorbing boundaries ($h = 0$ or $h = 1$) is reached. The survivor species will continue to fluctuate around its own steady state value.

Like before, the proportion of cells that become fixed on a certain allele is observed to be the same as the initial allele frequency. However, when w becomes fixed its steady state value is N_{opt}, whereas a fully mutant cell will have copy numbers fluctuating around αN_{opt}. If the simulation starts in steady state with the initial frequencies both 0.5 (meaning $w_0 = m_0 = \frac{N_{\text{opt}}\alpha}{1+\alpha}$), then eventually 50% of cells will be (on average) in state $(N_{\text{opt}}, 0)$ and 50% in state $(0, \alpha N_{\text{opt}})$. This means that for long times $\langle w \rangle = N_{\text{opt}}/2$ and $\langle m \rangle = \alpha N_{\text{opt}}/2$, i.e. $\langle w \rangle$ and $\langle m \rangle$ decrease and increase over time, respectively. Therefore, the heteroplasmy of the tissue as a whole ($\frac{\sum_i m_i}{\sum_i m_i + \sum_i w_i}$ where i denotes the ith cell) increases. Note, however, that it does not follow that the mean heteroplasmy *per cell*, $\frac{1}{n}\sum_i h_i$, increases with time.

A larger value of α means that $\langle m \rangle$ approaches a larger value, which seems to be disadvantageous. The benefit is, however, that w will deviate from its desired value N_{opt} more slowly as h increases. More precisely, $w_s/N_{\text{opt}} = (1-h)/(1-(1-1/\alpha)h)$, and a high α can therefore be interpreted as an attempt of the cell to keep the wildtype population near its optimal value. The consequence of a high α is an effective heteroplasmy threshold (Fig. 1) and this simple model could therefore be an explanation of the experimentally observed threshold effect [21–24]. A generalization of the model described in Eq. (4), in which the mutants were allowed to contribute to the feedback as well [i.e. $C = C(w, m)$], was described deterministically [76]. The more the mutants contribute to the feedback, the less pronounced the threshold effect seen in Fig. 1.

3.2 *In Silico Models of mtDNA Dynamics and Mitochondrial Dynamics*

In many models, mtDNA molecules are assumed to be well-mixed and each mtDNA has a given probability per unit time of being replicated and degraded. The occurrence of mitophagy events, events involving the degradation of a whole mitochondrion, means that all the mtDNA molecules within are simultaneously degraded. In this case, it becomes important to know which mtDNA resides in which mitochondrion. Moreover, mitochondria can only fuse with others when they are sufficiently close, meaning that spatial positions start to play a role. The possible roles of fission and fusion and the networks of mitochondria that are produced are unclear [8]. Various models describe fusion and fission dynamics [79–82], some of which include spatial effects [80, 82]. A brief overview of their results is given here.

The model in [82] incorporates random mtDNA turnover, fusion and fission of mitochondria, and spatial effects. MtDNA turnover was assumed to consist of (1) replication events of individual mtDNAs (with possible state feedback), and (2) degradation of whole mitochondria. The cell was divided into 16 compartments, and fusion only occurred between mitochondrial pairs in the same or in adjacent compartments. Upon fission, mtDNAs were binomially partitioned into daughter mitochondria which were themselves placed in their own compartment or an adjacent one. Some of these dynamics are described by the following Poisson processes:

$$
\begin{aligned}
(w,m) &\xrightarrow{a_R(w,m)} && (w+1,m)\\
(w,m) &\xrightarrow{a_R(w,m)} && (w,m+1)\\
(w,m) &\xrightarrow{a_D N_{\text{mito}}} && \emptyset\\
(w_1,m_1)+(w_2,m_2) &\xrightarrow{a_{fus}} && (w_1,m_1|w_2,m_2)\\
(w_1,m_1|w_2,m_2|w_3,m_3) &\xrightarrow{a_{\text{fis}}(w,m)} && (w_1',m_1')+(w_2',m_2'|w_3,m_3)\\
& && \text{with } w_1'+w_2'=w_1+w_2\\
& && \text{and } m_1'+m_2'=m_1+m_2
\end{aligned}
\tag{6}
$$

where (w,m) represents a single mitochondrion with w wildtype and m mutant mtDNA molecules, N_{mito} is the total number of mitochondria in the cell, $a_R(w,m)$ represents the replication rate with feedback ensuring upregulation of mtDNA copy number as heteroplasmy increases, and $(w_1,m_1|w_2,m_2)$ represents a fused mitochondrion. The third equation represents a mitophagy event in which an entire mitochondrion is being degraded, the last two equations give examples of fusion and fission events (any number of mitochondria can be fused together). The fission

propensity $a_{\text{fis}}(w, m)$ was assumed to increase with the size of the mitochondrion, i.e. its total mtDNA copy number. Stochastic Gillespie simulations [83] were used to model the system. Among the conclusions were the following: (1) faster fusion–fission dynamics results in a better mixing of mtDNAs, (2) slower mtDNA mixing increases the heteroplasmy variance between cells and speeds up the process of clonal expansion, (3) including replication feedback [similar to Eq. (4)] can lower the fraction of cells with clonally expanded mutants, but this effect is lessened with low fusion–fission rates.

The model described above was extended in [84] to account for the experimentally observed selectivity of mitochondrial fusion and mitophagy. Briefly, the higher the fraction of mutants inside a mitochondrion, the less likely it is to fuse and the more likely it is to be degraded. As in [82], higher rates of fusion and fission led to an increased heteroplasmy variance, but this was only beneficial when mitophagy and fusion were sufficiently selective, allowing for mitochondria with high h to be efficiently removed from the population. A decline of the selectivity of fusion and mitophagy with age can be a reason why mutants expand, and in this case a lower fusion–fission rate is actually beneficial [84].

In some other models the focus is not on mtDNA dynamics, but on the fusion and fission events themselves and how they affect the health of the overall mitochondrial population. It was assumed that mitochondria contain a discrete set of units (referred to as health units, hereditary units, or quality units) that can be exchanged during fusion–fission events and undergo damage over time [79–81]. Less healthy mitochondria (mitochondria with a low membrane potential) are less likely to fuse [85], and usually they are assumed to have a higher degradation rate. Because of this, higher rates of fusion and fission tend to lead to more healthy mitochondria [79–81]. When damaged mitochondria are able to spread their dysfunctions in an infection-like manner, lower fusion–fission rates were found to be beneficial [81].

Various other stochastic models of mitochondria were constructed, which are briefly summarized here: (1) the effect of a shorter mutant replication time was modelled using both DDEs (delayed differential equations) and stochastic simulations [68]. The authors concluded that faster mutant replication is highly unlikely to be the cause of clonal expansion; (2) the role of transcription rates in providing a replication advantage for mutants was investigated [36]. This model is discussed in more detail in Sect. 4; (3) Random drift was found to be sufficient to explain mutant loads in human tumors [44]; (4) a model simulating mtDNA segregation in hematopoietic stem cells found evidence for selection against pathogenic mutations [86]; and (5) a model was developed to investigate the dynamics of the network arising from fusion–fission events, the investigation of its equilibrium configurations in both deterministic and stochastic settings leading to the finding of the existence of percolation phase transition in the mitochondrial reticulum [72].

3.3 A General Approach to Investigating the Nuclear Control of mtDNA Dynamics

Recently, a general bottom-up theory has been produced to describe mtDNA dynamics in single cells [78]. The full model includes mtDNA turnover with (1) arbitrary copy number feedback control on replication and degradation rates, (2) cell divisions, (3) *de novo* mutations and replication errors, and (4) a possible selective advantage for mutant mtDNA molecules. Denoting a state with w wildtype and m mutant molecules by (w, m), the dynamics are described by the following set of Poisson processes:

$$\begin{aligned}
(w,m) &\xrightarrow{\epsilon_1+(1+\epsilon_2)w\lambda(w,m)\equiv f_1(w,m)} && (w+1,m)\\
(w,m) &\xrightarrow{\epsilon_3+(1+\epsilon_4)m\lambda(w,m)\equiv f_2(w,m)} && (w,m+1)\\
(w,m) &\xrightarrow{\epsilon_5+(1+\epsilon_6)w\nu(w,m)\equiv f_3(w,m)} && (w-1,m)\\
(w,m) &\xrightarrow{\epsilon_7+(1+\epsilon_8)m\nu(w,m)\equiv f_4(w,m)} && (w,m-1)\\
(w,m) &\xrightarrow{w\mu_1\equiv f_5(w,m)} && (w-1,m+1)\\
(w,m) &\xrightarrow{w\mu_2\equiv f_6(w,m)} && (w,m+1)\\
(w,m) &\xrightarrow{w\mu_3\equiv f_7(w,m)} && (w-1,m+2)
\end{aligned} \tag{7}$$

where $\lambda(w, m)$ and $\nu(w, m)$ are the replication and degradation rates, respectively, ϵ_i indicate any possible selective advantage in replication and/or degradation for w or m; this advantage can be multiplicative (even-indexed ϵ_i) or additive (odd-indexed ϵ_i), and μ_i indicate possible mutation processes; spontaneous mutations (μ_1), replication errors with the original molecule remaining intact (μ_2), and replication errors in which both the original and replicated molecule become mutated (μ_3). The rates of the reactions are given by f_j with $j = 1, \ldots, R$ where $R = 7$ is the total number of reactions.

The stoichiometry matrix of this system is given by

$$S_{ij} = \begin{pmatrix} 1 & 0 & -1 & 0 & -1 & 0 & -1 \\ 0 & 1 & 0 & -1 & 1 & 1 & 2 \end{pmatrix} \tag{8}$$

with the index j representing the different reactions given in (7), and $i = 1, \ldots, N$ denoting the different species (here, $i = 1$ corresponds to w and $i = 2$ to m, but the method can be readily extended to deal with more than 2 mtDNA species). Denoting $P_{w,m}(t)$ as the probability of observing the system in state (w, m) at time t, the system is described by the following master equation:

$$\begin{aligned}\frac{\partial P_{w,m}}{\partial t} &= \sum_{j=1}^{R}\left(\prod_{i=1}^{N}\mathbb{E}^{-S_{ij}}-1\right)f_j(w,m)P_{w,m}\\ &= \left(\mathbb{E}^{-S_{11}}\mathbb{E}^{-S_{21}}-1\right)f_1(w,m)P_{w,m}+\cdots+\left(\mathbb{E}^{-S_{17}}\mathbb{E}^{-S_{27}}-1\right)f_7(w,m)P_{w,m}\\ &= f_1(w-1,m)P_{w-1,m}-f_1(w,m)P_{w,m}+\cdots\\ &\quad +f_7(w+1,m-2)P_{w+1,m-2}-f_7(w,m)P_{w,m}\end{aligned} \tag{9}$$

where $\mathbb{E}^{-S_{ij}}$ is a raising and lowering operator.[1] For non-linear $f_j(w,m)$ this master equation is generally analytically intractable, but can be approximated by a van Kampen system size expansion [87]. The system size expansion treats w and m as the sum of a deterministic component (ϕ) and a fluctuating stochastic component (ξ), scaled by powers of the system size (Ω):

$$\begin{aligned}w &= \phi_w\Omega+\xi_w\Omega^{1/2}\\ m &= \phi_m\Omega+\xi_m\Omega^{1/2}\end{aligned} \tag{10}$$

All quantities in Eq. (9) are then written in terms of Ω, ϕ_i, and ξ_i, and equal powers of Ω are collected. The largest terms, proportional to $\Omega^{1/2}$, form the macroscopic rate equations, i.e. they describe the deterministic behaviour of the mean quantities in the system. Next, terms of order Ω^0, known as the linear noise approximaton (LNA), give a linear Fokker–Plank equation

$$\frac{\partial\Pi(\xi,t)}{\partial t}=\sum_{i,j=1}^{N}A_{ij}\frac{\partial\left(\xi_j\Pi\right)}{\partial\xi_i}+\frac{1}{2}\sum_{i,j=1}^{N}B_{ij}\frac{\partial^2\Pi}{\partial\xi_i\partial\xi_j} \tag{11}$$

with A_{ij},B_{ij} given by $A_{ij}=\sum_{k=1}^{R}S_{ik}\frac{\partial f_k}{\partial\phi_j}$ and $B_{ij}=\sum_{k=1}^{R}S_{ik}S_{jk}f_k$. Evolution equations for the moments of $\xi_w(t),\xi_m(t)$ [and, correspondingly, the moments of $w(t)$ and $m(t)$] can be extracted from this Fokker–Planck equation, forming a set of coupled ODEs.

For non-linear functions $\lambda(w,m)$ and $\nu(w,m)$ these coupled moment equations cannot be solved analytically, though they can be solved numerically. To make analytical progress, the replication and degradation rates can be linearized around their steady state values (w_{ss},m_{ss}) [78], i.e.

$$\begin{aligned}\lambda(w,m) &\approx \lambda(w_{ss},m_{ss})+\beta_w(w-w_{ss})+\beta_m(m-m_{ss})\\ \nu(w,m) &\approx \nu(w_{ss},m_{ss})+\delta_w(w-w_{ss})+\delta_m(m-m_{ss})\end{aligned} \tag{12}$$

[1] $\mathbb{E}^{-S_{ij}}$ removes S_{ij} from every occurrence of species i to its right, for example, $\mathbb{E}^{-S_{11}}f_1(w,m)P_{w,m}=f_1(w-1,m)P_{w-1,m}$ and $\prod_{i=1}^{N}\mathbb{E}^{-S_{i7}}=\mathbb{E}^{-S_{17}}\mathbb{E}^{-S_{27}}f_7(w,m)P_{w,m}=f_7(w+1,m-2)P_{w+1,m-2}$.

where $\beta_j = \partial_j \lambda(w,m)$ and $\delta_j = \partial_j \nu(w,m)$ with $j = w, m$.

Using this linearized system, full solutions for the means and variances of w and m over time were provided by the authors for a simplified version of Eq. (7). The existence of steady state values depends on the eigenvalues of the system's Jacobian matrix. One of the eigenvalues of the simplified system is zero, and imposing the conditions $\lambda(w_{ss}, m_{ss}) = \nu(w_{ss}, m_{ss})$, $\beta_w, \beta_m < 0$ and $\delta_w, \delta_m > 0$ ensures that the other eigenvalue is negative. This gives rise to a line of steady state values for w and m, and for timescales on which the LNA is valid the wildtype and mutant steady states are roughly constant.

Starting in steady state with noiseless initial conditions, the solutions can be written in the simple form:

$$\begin{aligned}
\langle w \rangle &= w_{ss} \\
\langle m \rangle &= m_{ss} \\
\langle w^2 \rangle &= F_1^{\text{decay}}(t) + \theta_1 t + \phi_1 \\
\langle wm \rangle &= F_2^{\text{decay}}(t) + \theta_2 t + \phi_2 \\
\langle m^2 \rangle &= F_3^{\text{decay}}(t) + \theta_3 t + \phi_3 \\
\langle h^2 \rangle' &\equiv \frac{\langle h^2 \rangle}{\langle h \rangle (1 - \langle h \rangle)} = \frac{2\lambda(w_{ss}, m_{ss})t}{w_{ss} + m_{ss}}
\end{aligned} \tag{13}$$

where F_i^{decay} are transient functions that die out exponentially with time, $\langle h^2 \rangle'$ is the normalized heteroplasmy variance (the quantity typically reported in experimental studies), $\langle h \rangle = \langle \frac{m}{w+m} \rangle$ is the expected heteroplasmy value (which was approximated by $\frac{\langle m \rangle}{\langle w \rangle + \langle m \rangle}$), and the constants θ_i and ϕ_i are functions only of (1) the difference between replication and degradation rates, (2) steady state copy numbers w_{ss} and m_{ss}, and (3) the turnover rate in steady state $\lambda(w_{ss}, m_{ss})$ $(= \nu(w_{ss}, m_{ss}))$. The structure of the above solutions is such that at nonzero w and m, at most one of the θ_i can be zero, and θ_1 and θ_3 are non-negative [78].

Several conclusions can be drawn from this linearized system when the assumptions underlying this derivation hold (see below): (1) the variance of at least one species (w or m) increases linearly with time, (2) heteroplasmy variance increases linearly with time with a rate depending only on steady state copy numbers and the timescale of random turnover. This last observation means that the rate of increase of $\langle h^2 \rangle$ does not depend on the specifics of the control mechanism applied (i.e. the specific forms of $\lambda(w,m)$ and $\nu(w,m)$), meaning that different control mechanisms lead to similar trends in heteroplasmy variance.

For arbitrary functions $\lambda(w,m)$, $\nu(w,m)$ and nonzero ϵ_i, μ_i in Eq. (7), the coupled ODEs provided by the system size expansion can be solved numerically, which allows one to study behaviours away from steady state. Various specific control mechanisms are investigated in [78], including the relaxed replication control given in Eq. (4). Numerical solutions generally agree well with stochastic simulations,

meaning that, when the LNA is valid (see below), variability arising from selection and mutation under *any* control mechanism can be characterized without requiring stochastic simulations.

It is further shown that with nonzero ϵ_8 and μ_2 (i.e. replication errors can occur and mutants are selectively degraded), mutants are only successfully cleared over time when $(1 + \epsilon_8) \gg \mu_2$ i.e. when selection is sufficiently strong to overcome the increase in m through mutations. Depending on the specific control mechanism used, the wildtype variance can significantly increase as the mutants are cleared.

All the above conclusions are based on the validity of the LNA. Behaviours of w and m start to deviate from the LNA at long times, or if extinction of one of the species (or both) becomes non-negligible. The values of w_{ss} and m_{ss}, which are roughly constant for short times, will start to change at longer times. The steady state of either species can increase or decrease over time and, depending on the details of the feedback functions, will either reach zero, settle down to a constant value, or increase unboundedly. When fixation occurs, the means and variances obtained through the LNA are underestimations of the actual means and variances, while heteroplasmy variance is overestimated by the LNA. Also, for general $\lambda(w, m)$ and $\nu(w, m)$, the transition rates between different states may be highly non-linear, making the LNA less accurate. Higher order correction terms in the system size expansion can be included to improve solutions, and a discussion on the accuracy of the LNA and its higher order terms is given in [88].

4 mtDNA and Ageing

As mentioned in the introduction, mtDNA mutations accumulate with age in any healthy individual. Recent reviews on the relationship between mtDNA mutations and ageing can be found in [89–91]. The mitochondrial theory of ageing proposes that accumulation of mitochondrial damage is the cause of ageing in humans and animals, but whether this causal relationship actually exists is still debated [92]. Here, various models are discussed which describe the accumulation of mutations through the process of random replication, degradation, and/or segregation of mtDNA molecules in cells.

Because of the random fluctuations in both wildtype and mutant mtDNA molecules in a cell, a mutant species can become homoplasmic purely by chance, without experiencing any selective advantage. This idea was modelled in non-dividing cells [43]. Cells were simulated over a human lifetime using the model described in Eq. (2), with the addition that every time an mtDNA replicates a mutation can occur with probability P_{mut}. Every mutation event is assumed to result in a new mutation, meaning that cells will acquire a variety of different mutations over time. It was shown that $P_{\text{mut}} = 5 \times 10^{-5}$ is sufficient to obtain 5–10% of cells with $h > 0.6$ after 100 years. Moreover, the majority (>80%) of mutated mtDNAs in the cell were the same, agreeing with the experimental observations of a single clonally expanded mutant (as opposed to many different mutations). The model that includes proliferation of pathological mutants [Eq. (4)] was also used,

but changing proliferation (i.e. changing the parameter α) had no significant effect on the accumulation of high heteroplasmy cells. The results they obtained suggest that random drift alone can indeed lead to clonal expansion on the scale of a human lifetime. It has further been hypothesized that random drift can also account for the accumulation of mutations seen in cancer and mitochondrial diseases [44, 93].

However, it has been argued that this model of mutant accumulation purely by chance, without any selective advantage for mutants, cannot explain clonal expansion in short-lived animals [41] such as rats and mice which have an average lifespan of only 3 years. For mutants to have expanded in 5–15% of all cells in such a short time, a much higher P_{mut} is required (7.6×10^{-3} vs 5×10^{-5} on a human lifetime). The problem with requiring such a high mutation rate is that the number of different mutations that are found in the cells at the end of the simulations is very high. On average, more than 30 types of mutants were present in each cell with $h > 0.6$ after 3 years. Moreover, the most frequent mutant in these cells represented less than 20% of all mtDNAs, meaning that experimentally several different mutants should be observed in these high heteroplasmy cells, which is not the case. This suggests that other mechanisms need to be evoked to explain clonal expansion in these short-lived animals.

Because some kind of replicative advantage seems to be required, several models were constructed to incorporate these advantages. In [68] a shorter replication time for mutants was used, but as mentioned in the previous section this particular mechanism is unlikely to explain clonal expansion. In [36], another mechanism was proposed. In order for mtDNA to be able to replicate, it needs to be transcribed as well, which results in the production of proteins. A negative feedback loop is assumed to exist, i.e. transcription rates of mtDNA drop if protein levels are high enough. Deletion mutants miss large parts of their DNA, so some proteins are not produced at all, meaning that the negative feedback on transcription rate never occurs, resulting in higher replication levels. An ODE model was constructed describing the dynamics of w, m and the level of ATP. The presence of mtDNA molecules is assumed to consume some ATP due to the requirements of producing and maintaining mitochondrial machinery (likewise for other cellular processes), and w is assumed to produce ATP. If ATP levels are high, replication of w and m is low. The higher replication rate of m (assumed to be 50% higher than that of w) makes them increase exponentially, eventually leading to system collapse through ATP exhaustion. The ODE model manages to explain the accumulation of mutants for short-lived animals, and levels of w and m after collapse agree with experimental observations. Next a stochastic model based on the ODE reactions was developed. The values for P_{mut} were adjusted such that at the end of the simulations (the final simulation time ranging from 3 to 80 years to model different organisms) $10 \pm 0.5\%$ of cells have $h > 0.6$. The model does not require very high mutation rates in short-lived animals, and also predicts that the average number of different mtDNAs present at the end of the simulations is around 1 [36]. To obtain the desired $10 \pm 0.5\%$ of cells with $h > 0.6$ in humans, very low mutation probabilities ($P_{\text{mut}} \sim 10^{-7}$) are required. This seems at odds with the finding that P_{mut} is highly likely to lie between 10^{-4} and 10^{-5} in substantia nigra neurons [60], though mtDNA mutation rates may depend on cell type.

5 mtDNA and Development

Unfertilized human eggs contain on the order of 10^5 copies of mtDNA, some of which may be mutated. After fertilization, the egg starts dividing, and with each division the mtDNAs are stochastically partitioned between the daughter cells. By chance, some daughter cells will inherit more mutants than others, introducing a variance in heteroplasmy across the population of cells (see, e.g., [94]). This allows for the elimination of cells with high mutant loads (as they can become dysfunctional and initiate cell death), while purely wildtype cells and cells with low h survive. An increased heteroplasmy variance therefore provides a mechanism of filtering out mutant mtDNA molecules to reduce mutant load (and thus increase the health) of the offspring. Even though the starting mtDNA copy number of the egg is very high, the copy number per cell falls drastically in early development because of rapid cell divisions with little replication of mtDNA. This fall in copy number per cell further increases the stochasticity and thus the heteroplasmy variance between cells, and is termed the mitochondrial bottleneck.

The exact mechanism by which heteroplasmy variance increases, however, is highly debated. It might be that random drift in copy numbers between cell divisions, stochastic partitioning at cell divisions, or both, is sufficient to explain the observed heteroplasmy variance [59, 78, 95–99]. However, other studies show a less pronounced decrease in mtDNA copy number per cell [100]. Additional bottlenecking mechanisms were suggested, such as the clustering of mtDNAs during cell division (increasing the stochasticity upon division) [100], or restricting the ability to replicate to only a subpopulation of mtDNAs [101]. Recently, a general model was developed which was able to reproduce all of these mechanisms and, importantly, provides a statistical framework to compare them given experimental observations [102]. Approximate Bayesian Computation (ABC) [103, 104] was used to infer the statistical support for each of the three mentioned mechanisms. Overall, the most support is found for mechanism involving a combination of random mtDNA turnover and binomial partitioning at cell division.

Several models have been developed to describe the behaviour of mtDNA heteroplasmy through development. Some of these are summarized here, in order of increasing complexity. The first model, originating from population dynamics and termed the Wright formula [105], describes the heteroplasmy mean and variance of a population over any number of cell divisions with binomial partitioning rules. A more detailed model was developed using the Kimura distribution [106] (also originally used in population genetics). The Kimura distribution allows a description of the entire heteroplasmy probability distribution after any number of cell divisions. Recently, a model was developed that includes mtDNA dynamics in between cell divisions [102, 107], providing analytical expressions of the heteroplasmy probability generating function after any number of cell divisions with birth-immigration-death dynamics in between divisions.

5.1 The Wright Formula of Partitioning Variance

Population genetics studies the frequency and interaction of alleles and genes in populations. Genetic drift in allele frequencies arises because the alleles of the offspring are randomly sampled from those of the parents. If random genetic drift is the only force acting on an allele then, after n generations, the variance in allele frequency across a population is given by the Wright formula:

$$V_n = p(1-p)\left(1-\left(1-\frac{1}{N_{\text{eff}}}\right)^n\right) \tag{14}$$

where p and $(1-p)$ are the initial allele frequencies, and N_{eff} is an effective population size [105]. The mean allele frequency is assumed to be equal to the initial allele frequency.

To apply this theory to an mtDNA population, the Wright formula can be interpreted as describing the variance in mutant allele frequency, i.e. heteroplasmy, over n cell divisions. At cell division, each daughter cell obtains N_{eff} mtDNA molecules which are randomly sampled with replacement from the mother cell. This approach has been used in various studies of mtDNA heteroplasmy [96, 97, 99, 108, 109]. Some of the conclusions drawn from these studies must be taken with care, as it was shown that including a principled description of the uncertainty arising from sampling small populations might change the interpretation of the data [110].

Applied to mtDNA dynamics, the formula has several limitations. Firstly, the parameter N_{eff} is hard to interpret and does not correspond to any biological entity in the cell [111]. Secondly, knowing only the mean and variance of the heteroplasmy distribution may not be very informative. Finally, the Wright formula ignores stochastic effects resulting from random turnover of mtDNA between cell divisions. As is shown in the next section, including mtDNA turnover leads to a correction of the formula.

5.1.1 The Inclusion of mtDNA Turnover Between Cell Divisions

The model described earlier in Sect. 3.3 was used to adjust the Wright formula to include turnover of mtDNA molecules in between cell divisions [78]. The expected heteroplasmy variance given in Eq. (13) describes the approximate steady state variance in h arising from random mtDNA turnover through birth–death dynamics. This formula does not include cell divisions, and the denominator of the equation, $w_{ss} + m_{ss}$, thus gives the expected population size without the inclusion of cell divisions. If n is the population size immediately after a cell divides, then in order to maintain a constant average population size n has to increase to $2n$ before the next cell division. The expected population size is then roughly given by $\frac{3}{2}n$. Here, no reference to a particular feedback mechanism was made, but if more knowledge about the feedback mechanism is present, then the expected population size can

be tailored more appropriately. In the general case, the approximate expected heteroplasmy variance through random mtDNA turnover *and* cellular turnover is then given by:

$$\begin{aligned}\langle h^2\rangle' &= \frac{2\lambda t}{n_{\text{eff}}} \\ &= \frac{4t}{3n\tau}\end{aligned} \tag{15}$$

where n is the number of mtDNA molecules immediately after cell division, and τ is the timescale of mtDNA degradation (assuming that $\lambda = \nu = 1/\tau$, i.e. constant birth and death rates that are balanced to keep a constant average population size). This leads to the 'turnover adjusted Wright formula' proposed in Ref. [78]:

$$\langle h^2\rangle' = 1 - \left(1 - \frac{1}{2n}\right)^g + \frac{4t}{3n\tau} \tag{16}$$

where g is the number of generations (cell divisions) that have occurred, t is the amount of time that has expired since an initial state with $\langle h^2\rangle' = 0$, and τ is the timescale of mtDNA degradation. This adjusted Wright formula now includes random turnover of mtDNA and is written in terms of observable quantities, namely n (the number of mtDNA immediately after cell division), g (the number of cell divisions), and τ (the timescale of mtDNA turnover). While these models are useful in some circumstances, a more detailed approach has benefits as well and was developed by Kimura [106] as described in the next section. Equation (16) was tested against stochastic simulations and shown to be an improvement on the original formula [78].

5.2 The Kimura Distribution

To predict the entire heteroplasmy distribution after any number of cell divisions, the theory developed by Motoo Kimura [106] was applied to mtDNA segregation [112]. The Kimura distribution describes gene frequency distributions under random genetic drift. It is assumed that there is no selection and that there are no *de novo* mutations. A lack of *de novo* mutations means that for long times, the heteroplasmy in cells will settle down on either fully wildtype or fully mutant, as these are the only two absorbing states. According to the Kimura model, the total probability distribution for a particular allele (e.g. mutant mtDNAs) consists of three probability distributions: (1) the probability $f(0,t)$ for having lost the allele in generation t, (2) the probability $f(1,t)$ for having fixed on that allele, and (3) the probability distribution $\phi(x,t)$ giving the probability of observing the allele at frequency x in generation t:

$$f(0,t) = (1-p_0) + \sum_{i=1}^{\infty}(2i+1)p_0(1-p_0)(-1)^i \times F(1-i,i+2,2,1-p_0)e^{-i(i+1)/(2N_{\text{eff}})t} \tag{17}$$

$$\phi(x,t) = \sum_{i=1}^{\infty} i(i+1)(2i+1)p_0(1-p_0) \times F(1-i,i+2,2,x)F(1-i,i+2,2,p_0)e^{-i(i+1)/(2N_{\text{eff}})t} \tag{18}$$

$$f(1,t) = p_0 + \sum_{i=1}^{\infty}(2i+1)p_0(1-p_0)(-1)^i \times F(1-i,i+2,2,p_0)e^{-i(i+1)/(2N_{\text{eff}})t} \tag{19}$$

where $F(a,b,c,d)$ is the hypergeometric function and p_0 is the initial allele frequency [106]. The variance of the Kimura distribution is the same as the variance described by the Wright formula in Eq. (14). Interpreting the Kimura model for mtDNA segregation means that p_0 is the initial heteroplasmy, and $f(0,t)$, $f(1,t)$, and $\phi(x,t)$ are the probabilities of observing $h = 0$, $h = 1$, and $h = x$ (with $x \neq 0, 1$) after t cell divisions, respectively. These equations were used to describe heteroplasmy data from human, mouse, and *Drosophila* [112]. Overall, the Kimura distribution provides a good description of experimental data [112]. In [112] only the heteroplasmy mean and variance were matched to data, but more detailed fits are possible using an explicit likelihood function. Some experimentally reported increases in heteroplasmy variance become hard to defend if standard errors of the variances are taken into account, assuming that the heteroplasmy variance data is sampled from a Kimura distribution [110].

An alternative way to obtain the full heteroplasmy distribution is by using stochastic simulations. The advantage of simulations is that *de novo* mutations and selection mechanisms can be easily included, though they do not provide an explicit analytical distribution as is given in Eqs. (17)–(19).

5.3 *Analytical Descriptions of Random Turnover Combined with Cell Divisions*

To describe the dynamics of mtDNA molecules over time through cycles of cell divisions, both the turnover within a cell cycle and the partitioning at cell divisions have to be taken into account. An analytical description of these dynamics are described in a recent model which follows the probability distribution of an agent (e.g. an mtDNA molecule) over time. The dynamics of the agents is assumed to arise from a combination of: (1) random turnover of agents between cell divisions according to a birth–death-immigration (BID) model, and (2) stochastic partitioning

at cell division [107]. Several possible partitioning schemes are considered and analytic results are demonstrated for two important examples: binomial partitioning and subtractive partitioning. In subtractive partitioning a small number of agents are transferred to a small bud and the larger cell that is left over is tracked in the next generation, a model which is appropriate in various organisms such as budding yeast. A similar approach is taken in [102], with birth–death dynamics in between cell divisions. Here, a summary of the approach taken in [102] is given and some conclusions of the models in both [102] and [107] are discussed.

5.3.1 Within Cell Cycle Dynamics

Within a cell cycle, the time evolution of the probability distribution $P_m(t)$ of observing m agents at time t, according to a birth death model is given by the following master equation:

$$\frac{\partial P_m(t)}{\partial t} = \nu(m+1)P_{m+1}(t) + \lambda(m-1)P_{m-1}(t) - (\nu+\lambda)mP_m(t) \tag{20}$$

where λ and ν are the birth and death rates, respectively. The initial condition is assumed to be $P_m(0) = \delta_{m,m_0}$. A solution can be obtained by solving for the probability generating function $G(z,t) = \sum_{m=0}^{\infty} z^m P_m(t)$ [113]. Knowing the generating function $G(z,t)$ is equivalent to knowing the probability distribution because all the moments from the distribution can be derived from its derivatives. The generating function of the birth–death model satisfies

$$\begin{aligned}\frac{\partial G(z,t)}{\partial t} &= \sum_{m=0}^{\infty} z^m \frac{\partial P_m(t)}{\partial t} \\ &= \Big(\nu(1-z) + \lambda(z^2-z)\Big)\frac{\partial G(z,t)}{\partial z}\end{aligned} \tag{21}$$

with $G(z,0) = z^{m_0}$, which can be solved [102, 107] to give

$$\begin{aligned}G(z,t|m_0) &= \left(\frac{(z-1)\nu e^{(\lambda-\nu)t} - \lambda z + \nu}{(z-1)\lambda e^{(\lambda-\nu)t} - \lambda z + \nu}\right)^{m_0} \\ &\equiv \left(\frac{Az+B}{Cz+D}\right)^{m_0} \\ &\equiv g(z,t)^{m_0}\end{aligned} \tag{22}$$

where $A = \nu l - \lambda$, $B = \nu - \nu l$, $C = \lambda l - \lambda$, and $D = \nu - \lambda l$ with $l = e^{(\lambda-\nu)t}$.

When birth and death are balanced ($\lambda = \nu$) Eq. (22) can be rewritten to give

$$G_{\lambda=\nu}(z,t|m_0) = \left(\frac{\nu t - z - \nu t}{\nu t z - 1 - \nu t}\right)^{m_0} \tag{23}$$

which is obtained by writing $\lambda = \nu + \epsilon$ in Eq. (22) and taking the limit $\epsilon \to 0$.

5.3.2 Agent Partitioning at Cell Division

The overall generating function of the process containing both cell divisions and birth–death dynamics between these divisions can be written [102, 107] in a similar form to Eq. (22), i.e.

$$g_{\text{div}}(z,t,n) = \left(\frac{A_{\text{div}} z + B_{\text{div}}}{C_{\text{div}} z + D_{\text{div}}}\right) \tag{24}$$

with

$$\begin{aligned}
A_{\text{div}} &= 2^n \lambda (l + l' - 2) - l^n l' (\lambda + \nu(l-2)) \\
B_{\text{div}} &= l^n l' (\lambda + \nu(l-2)) - 2^n (\lambda l' + \nu(l-2)) \\
C_{\text{div}} &= -\lambda l^n l' (l-1) + 2^n \lambda (l + l' - 2) \\
D_{\text{div}} &= \lambda l^n l' (l-1) - 2^n (\lambda l' + \nu(l-2))
\end{aligned} \tag{25}$$

where n is the number of cell divisions, λ and ν are the birth and death rates of the mtDNA dynamics in between each of these divisions, $l' = e^{(\lambda-\nu)t}$, and $l = e^{(\lambda-\nu)\tau}$ with τ the length of the cell cycle (which is here assumed to be equal for all cell cycles) [102, 107]. Equation (22) is a special case of Eq. (24) with $n \to 0$ and $\tau \to 0$.

5.3.3 Combined Overall Solution

To construct a generating function of the overall process with n_p phases, the generating functions of each phase have to be linked together in an appropriate way. Denoting the parameters describing the ith phase with index i (e.g. A_i corresponds to A in Eq. (25) with λ, ν, n replaced by λ_i, ν_i, n_i), the overall generating function is given by [102, 107]:

$$\begin{aligned}
g_{\text{overall}} &= \frac{A' z + B'}{C' z + D'} \\
G_{\text{overall}} &= g_{\text{overall}}^{m_0}
\end{aligned} \tag{26}$$

with

$$\begin{pmatrix} A' & B' \\ C' & D' \end{pmatrix} = \prod_{i=1}^{n_p} \begin{pmatrix} A_i & B_i \\ C_i & D_i \end{pmatrix} \tag{27}$$

From this overall generating function, means and variances (and, if necessary, the full probability distribution) of m at any time t throughout development can be calculated. By performing a similar analysis for the wildtype population w, heteroplasmy statistics can also be obtained. This allows for the evaluation of important quantities such as the mean and variance of $h(t)$ and the probability of crossing a certain heteroplasmy threshold (i.e. $P(h > h^*)$ for some h^*) throughout development, and the mutant fixation probability $P(m = 0, t)$.

One of the conclusions drawn from the model is that an increase in mitochondrial degradation increases heteroplasmy variance, and therefore increases the strength of selection to remove high heteroplasmy cells. This means that clinically increasing mitochondrial degradation may represent a way to reduce heteroplasmy levels in offspring. The more general model considered in [107] provides full, closed-form generating functions for several types of mtDNA dynamics, making it possible to extract all details of copy number distributions at any given time. The approach provides a way to explore the statistics of systems of mtDNA dynamics, with or without cell divisions, with arbitrarily changing population size. This formalism could also be applied to, for example, mtDNA dynamics in tumour cells.

6 Discussion

Mitochondrial dysfunctions and mutations are linked to many different diseases and it is important to understand how these dysfunctions arise, how they develop over time, and how they can be treated. Mathematical models are valuable tools to explore these questions from both the perspective of fundamental biology and of clinical strategies. In this review, the focus is on models of mtDNA dynamics and mitochondrial fusion–fission dynamics, the accumulation of mutant mtDNAs over time, and the role of mtDNA in ageing and development. Of particular importance is the time evolution of heteroplasmy values, since the proportion of cells exceeding a critical heteroplasmy threshold is related to disease severity. Numerous models, deterministic and stochastic, have been constructed describing mtDNA dynamics in different cells and over various timescales. Deterministic models typically describe mean behaviours of heteroplasmy alone; stochastic approaches are vital to describe the biomedically central structure of mtDNA distributions [78, 102, 107].

Existing models cover different mitochondrial aspects and use various approaches. At high copy numbers, mtDNA dynamics are well described by deterministic models (e.g. [33, 38, 39, 76]). The modelling of low copy numbers, fixation probabilities, and population variances requires the construction of

stochastic models. Both simulation-based models (e.g. [43, 44, 68, 80, 82, 84]), analytical models (e.g. [78, 102, 105–107]) and numerical approximations to analytical models (e.g. [78]) have been made, some of which include spatial effects (e.g. [80, 82, 84]) or mitochondrial fusion–fission dynamics (e.g. [79–82, 84]). The nuclear control of mtDNA copy number has been modelled in several ways, e.g. (1) a simple total copy number control [75]; (2) negative feedback control of replication rates dependent on the wildtype [75] or both wildtype and mutant species [76] using proportional selection; and (3) a more general negative feedback control for both random replication and degradation rates.

These mathematical approaches have led to many new discoveries and progress. Predictions made by the relaxed replication model have found experimental support [77, 114]. A recent model [115] increased the power of analysis of large-scale experiments to identify a potential issue with cutting-edge medical therapies: namely, that proliferative differences between different mtDNA types, as may arise in therapeutic contexts, can lead to amplification of potentially harmful mutant mtDNA. The evidence from this joint mathematical and experimental study influenced the UK HFEA's policy decisions on the implementation of these therapies [116]; the 'haplotype matching' approach that it advocated is now gaining support [117]. A model of the mitochondrial bottleneck [102] suggested approaches by which drugs can be used to modulate mtDNA during development and ameliorate disease inheritance, which is now being tested experimentally [118]. Additionally, this model provided clinically motivated strategies for optimally sampling embryos in preimplantation genetic diagnoses to address the inheritance of mtDNA diseases [102]. The many hypotheses existing on how exactly mutant mtDNA molecules expand over time can be tested with simulations and results indicate that (1) random mutant accumulation without any selection advantage has so far failed to explain clonal expansion in short-lived animals, (2) it is unlikely that a shorter mutant replication time causes clonal expansion [68], and (3) a higher mutant replication rate produces outcomes that are roughly consistent with various experimental data [36]. The ability of stochastic mathematical models to test, falsify, or confirm these hypotheses is extremely valuable because knowing how and why mutants accumulate allows us to clinically intervene in this process and potentially create a treatment for mitochondrial diseases.

Despite large amounts of progress, there are still open problems. Is there a causal relationship between mitochondrial damage and ageing? Mice accumulating mutations on a faster timescale ('mutator mice') show accelerated ageing-like phenotypes and shortened lifespan. However, this is only true for homozygous mice; heterozygous mice do not show ageing phenotypes despite having high mutation burdens [119]. How does the cell regulate its mitochondrial copy number? As noted in [78] it is experimentally hard to distinguish between different nuclear feedback mechanisms as distinct mechanisms can lead to very similar dynamics. The mechanisms by which mutant mtDNA molecules expand is still not fully understood, which is reflected by the large amount of hypotheses put forward. The recently developed model in [36] requires a very low mutation probability in human cells, which is not true for all cell types [60]. Moreover, it may not be able to

explain clonal expansion of certain point mutations. It is likely that preferential replication of mutants is the result of a combination of multiple mechanisms which are different for distinct mutations and cell types, making a general theoretical description very challenging. We anticipate the development of stochastic models for mtDNA populations to continue to produce scientific insights as the amount of experimental data characterizing this rich and medically important system increases in the future.

References

1. W. Martin, M. Mentel, The origin of mitochondria. Nat. Educ. **3**(9), 58 (2010)
2. R. Lill, B. Hoffmann, S. Molik, A.J. Pierik, N. Rietzschel, O. Stehling et al., The role of mitochondria in cellular iron–sulfur protein biogenesis and iron metabolism. Biochim. Biophys. Acta Mol. Cell Res. **1823**(9), 1491–1508 (2012)
3. C. Wang, R.J. Youle, The role of mitochondria in apoptosis. Annu. Rev. Genet. **43**, 95 (2009)
4. R. Rizzuto, D. De Stefani, A. Raffaello, C. Mammucari, Mitochondria as sensors and regulators of calcium signalling. Nat. Rev. Mol. Cell Biol. **13**(9), 566–578 (2012)
5. J. Aryaman, H. Hoitzing, J.P. Burgstaller, I.G. Johnston, N.S. Jones, Mitochondrial heterogeneity, metabolic scaling and cell death. BioEssays (2017). doi:10.1002/bies201700001
6. K. Mitra, C. Wunder, B. Roysam, G. Lin, J. Lippincott-Schwartz, A hyperfused mitochondrial state achieved at G1'S regulates cyclin E buildup and entry into S phase. Proc. Natl. Acad. Sci. U. S. A. **106**(29), 11960–11965 (2009). doi:10.1073/pnas.0904875106
7. M. Liesa, M. Palacín, A. Zorzano, Mitochondrial dynamics in mammalian health and disease. Physiol. Rev. **89**(3), 799–845 (2009). doi:10.1152/physrev.00030.2008
8. H. Hoitzing, I.G. Johnston, N.S. Jones, What is the function of mitochondrial networks? A theoretical assessment of hypotheses and proposal for future research. Bioessays **37**(6), 687–700 (2015)
9. H. Chen, S. Ren, C. Clish, M. Jain, V. Mootha, J.M. McCaffery et al., Titration of mitochondrial fusion rescues Mff-deficient cardiomyopathy. J. Cell Biol. **211**(4), 795–805 (2015)
10. V.S.V. Laar, S.B. Berman, Mitochondrial dynamics in Parkinson's disease. Exp. Neurol. **218**(2), 247–256 (2009). doi:http://dx.doi.org/10.1016/j.expneurol.2009.03.019
11. S. Grandemange, S. Herzig, J.C. Martinou, Mitochondrial dynamics and cancer. Semin. Cancer Biol. **19**(1), 50–56 (2009)
12. X. Zhu, G. Perry, M.A. Smith, X. Wang, Abnormal mitochondrial dynamics in the pathogenesis of Alzheimer's disease. J. Alzheimers Dis. **33**, S253–S262 (2013)
13. D.C. Chan, Fusion and fission: interlinked processes critical for mitochondrial health. Annu. Rev. Genet. **46**(1), 265–287 (2012) doi:10.1146/annurev-genet-110410-132529
14. H. Chen, D.C. Chan, Mitochondrial dynamics - fusion, fission, movement, and mitophagy - in neurodegenerative diseases. Hum. Mol. Genet. **18**(R2), R169–R176 (2009). doi:10.1093/hmg/ddp326
15. I.G. Johnston, B.P. Williams, Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. Cell Syst. **2**(2), 101–111 (2016)
16. D. Bogenhagen, D.A. Clayton, Mouse L cell mitochondrial DNA molecules are selected randomly for replication throughout the cell cycle. Cell **11**(4), 719–727 (1977)
17. L. Chatre, M. Ricchetti, Prevalent coordination of mitochondrial DNA transcription and initiation of replication with the cell cycle. Nucleic Acids Res. **41**(5), 3068–3078 (2013)
18. M. Alexeyev, I. Shokolenko, G. Wilson, S. LeDoux, The maintenance of mitochondrial DNA integrity - critical analysis and update. Cold Spring Harb. Perspect. Biol. **5**(5), a012641 (2013)
19. I.J. Holt, A. Reyes, Human mitochondrial DNA replication. Cold Spring Harb. Perspect. Biol. **4**(12), a012971 (2012)

20. G.S. Gorman, A.M. Schaefer, Y. Ng, N. Gomez, E.L. Blakely, C.L. Alston et al., Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. Ann. Neurol. **77**(5), 753–759 (2015)
21. G. Attardi, M. Yoneda, A. Chomyn, Complementation and segregation behavior of disease-causing mitochondrial DNA mutations in cellular model systems. Biochim. Biophys. Acta Mol. Basis Dis. **1271**(1), 241–248 (1995)
22. L. Boulet, G. Karpati, E. Shoubridge, Distribution and threshold expression of the tRNA (Lys) mutation in skeletal muscle of patients with myoclonic epilepsy and ragged-red fibers (MERRF). Am. J. Hum. Genet. **51**(6), 1187 (1992)
23. K. Nakada, K. Inoue, T. Ono, K. Isobe, A. Ogura, Y. Goto et al., Inter-mitochondrial complementation: mitochondria-specific system preventing mice from expression of disease phenotypes by mutant mtDNA. Nat. Med. **7**(8), 934–940 (2001)
24. C.T. Moraes, E.A. Schon, Detection and analysis of mitochondrial DNA and RNA in muscle by in situ hybridization and single-fiber PCR. Methods Enzymol. **264**, 522–540 (1996)
25. H.R. Elliott, D.C. Samuels, J.A. Eden, C.L. Relton, P.F. Chinnery, Pathogenic mitochondrial DNA mutations are common in the general population. Am. J. Hum. Genet. **83**(2), 254–260 (2008)
26. B.A. Payne, I.J. Wilson, P. Yu-Wai-Man, J. Coxhead, D. Deehan, R. Horvath et al., Universal heteroplasmy of human mitochondrial DNA. Hum. Mol. Genet. **22**(2), 384–390 (2013)
27. E.A. Schon, E. Bonilla, S. DiMauro, Mitochondrial DNA mutations and pathogenesis. J. Bioenerg. Biomembr. **29**(2), 131–149 (1997)
28. M. Bogliolo, A. Izzotti, S. De Flora, C. Carli, A. Abbondandolo, P. Degan, Detection of the4977 bp'mitochondrial DNA deletion in human atherosclerotic lesions. Mutagenesis **14**(1), 77–82 (1999)
29. Y. Michikawa, F. Mazzucchelli, N. Bresolin, G. Scarlato, G. Attardi, Aging-dependent large accumulation of point mutations in the human mtDNA control region for replication. Science **286**(5440), 774–779 (1999)
30. D.C. Wallace, Mitochondrial genetics: a paradigm for aging and degenerative diseases? Science **256**(5057), 628 1992
31. A. Kowald, E. Klipp, Mathematical models of mitochondrial aging and dynamics. Prog. Mol. Biol. Transl. Sci. **127**, 63–92 (2014)
32. A.D. De Grey, A proposed refinement of the mitochondrial free radical theory of aging. Bioessays **19**(2), 161–166 (1997)
33. A. Kowald, T.B. Kirkwood, Accumulation of defective mitochondria through delayed degradation of damaged organelles and its possible role in the ageing of post-mitotic and dividing cells. J. Theor. Biol. **202**(2), 145–160 (2000)
34. M. Yoneda, A. Chomyn, A. Martinuzzi, O. Hurko, G. Attardi, Marked replicative advantage of human mtDNA carrying a point mutation that causes the MELAS encephalomyopathy. Proc. Natl. Acad. Sci. **89**(23), 11164–11168 (1992)
35. E.A. Shoubridge, G. Karpati, K.E. Hastings, Deletion mutants are functionally dominant over wild-type mitochondrial genomes in skeletal muscle fiber segments in mitochondrial disease. Cell **62**(1), 43–49 (1990)
36. A. Kowald, T.B. Kirkwood, Transcription could be the key to the selection advantage of mitochondrial deletion mutants in aging. Proc. Natl. Acad. Sci. **111**(8), 2972–2977 (2014)
37. D. Harman, Free radical theory of aging: dietary implications. Am. J. Clin. Nutr. **25**(8), 839–843 (1972)
38. A. Kowald, T. Kirkwood, Mitochondrial mutations, cellular instability and ageing: modelling the population dynamics of mitochondria. Mutat. Res./DNAging **295**(3), 93–103 (1993)
39. A. Kowald, T. Kirkwood, A network theory of ageing: the interactions of defective mitochondria, aberrant proteins, free radicals and scavengers in the ageing process. Mutat. Res./DNAging **316**(5), 209–236 (1996)
40. C.B. Park, N.G. Larsson, Mitochondrial DNA mutations in disease and aging. J. Cell Biol. **193**(5), 809–818 (2011)

41. A. Kowald, T.B. Kirkwood, Mitochondrial mutations and aging: random drift is insufficient to explain the accumulation of mitochondrial deletion mutants in short-lived animals. Aging Cell **12**(4), 728–731 (2013)
42. I.J. Holt, D. Speijer, T.B. Kirkwood, The road to rack and ruin: selecting deleterious mitochondrial DNA variants. Philos. Trans. R. Soc. B **369**(1646), 20130451 (2014)
43. J. Elson, D. Samuels, D. Turnbull, P. Chinnery, Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. Am. J. Hum. Genet. **68**(3), 802–806 (2001)
44. H.A. Coller, K. Khrapko, N.D. Bodyak, E. Nekhaeva, P. Herrero-Jimenez, W.G. Thilly, High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. Nat. Genet. **28**(2), 147–150 (2001)
45. F.J. Miller, F.L. Rosenfeldt, C. Zhang, A.W. Linnane, P. Nagley, Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. Nucleic Acids Res. **31**(11), e61–e61 (2003)
46. P. Kaufmann, S. Shanske, M. Hirano, S. DiMauro, M.P. King, Y. Koga et al., Mitochondrial DNA and RNA processing in MELAS. Ann. Neurol. **40**(2), 172–180 (1996)
47. C.T. Moraes, E. Ricci, V. Petruzzella, S. Shanske, S. DiMauro, E. A. Schon et al., Molecular analysis of the muscle pathology associated with mitochondrial DNA deletions. Nat. Genet. **1**(5), 359–367 (1992)
48. M. Tokunaga, S. Mita, T. Murakami, T. Kumamoto, M. Uchino, I. Nonaka et al., Single muscle fiber analysis of mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes (MELAS). Ann. Neurol. **35**(4), 413–419 (1994)
49. J.M. Shoffner, M.T. Lott, A.M. Lezza, P. Seibel, S.W. Ballinger, D.C. Wallace, Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRNA Lys mutation. Cell **61**(6), 931–937 (1990)
50. Y. Goto, I. Nonaka, S. Horai, A mutation in the tRNALeu (UUR) gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. Nature **348**, 651–653 (1990)
51. N.W. Soong, D.R. Hinton, G. Cortopassi, N. Arnheim, Mosaicism for a specific somatic mitochondrial DNA mutation in adult human brain. Nat. Genet. **2**(4), 318–323 (1992)
52. M. Corral-Debrinski, T. Horton, M.T. Lott, J.M. Shoffner, M.F. Beal, D.C. Wallace, Mitochondrial DNA deletions in human brain: regional variability and increase with advanced age. Nat. Genet. **2**(4), 324–329 (1992)
53. N.G. Larsson, Somatic mitochondrial DNA mutations in mammalian aging. Annu. Rev. Biochem. **79**, 683–706 (2010)
54. E.J. Brierley, M.A. Johnson, R.N. Lightowlers, O.F. James, D.M. Turnbull, Role of mitochondrial DNA mutations in human aging: implications for the central nervous system and muscle. Ann. Neurol. **43**(2), 217–223 (1998)
55. J. Müller-Höcker, Cytochrome c oxidase deficient fibres in the limb muscle and diaphragm of man without muscular disease: an age-related alteration. J. Neurol. Sci. **100**(1), 14–21 (1990)
56. D. Cottrell, P. Ince, E. Blakely, M. Johnson, P. Chinnery, M. Hanna et al., Neuropathological and histochemical changes in a multiple mitochondrial DNA deletion disorder. J. Neuropathol. Exp. Neurol. **59**(7), 621–627 (2000)
57. Herbst A, Pak JW, McKenzie D, Bua E, Bassiouni M, Aiken JM. Accumulation of mitochondrial DNA deletion mutations in aged muscle fibers: evidence for a causal role in muscle fiber loss. J. Gerontol. A Biol. Sci. Med. Sci. **62**(3), 235–245 (2007)
58. Z. Cao, J. Wanagat, S.H. McKiernan, J.M. Aiken, Mitochondrial DNA deletion mutations are concomitant with ragged red regions of individual, aged muscle fibers: analysis by laser-capture microdissection. Nucleic Acids Res. **29**(21), 4502–4508 (2001)
59. S.K. Poovathingal, J. Gruber, B. Halliwell, R. Gunawan, Stochastic drift in mitochondrial DNA point mutations: a novel perspective ex silico. PLoS Comput Biol. **5**(11), e1000572 (2009)
60. D.A. Henderson, R.J. Boys, K.J. Krishnan, C. Lawless, D.J. Wilkinson, Bayesian emulation and calibration of a stochastic computer model of mitochondrial DNA deletions in substantia nigra neurons. J. Am. Stat. Assoc. **2014**, 76–87 (2012)

61. J.L. Hayashi, S. Ohta, A. Kikuchi, M. Takemitsu, Y. Goto, I. Nonaka, Introduction of disease-related mitochondrial DNA deletions into HeLa cells lacking mitochondrial DNA results in mitochondrial dysfunction. Proc. Natl. Acad. Sci. **88**(23), 10614–10618 (1991)
62. A. Chomyn, A. Martinuzzi, M. Yoneda, A. Daga, O. Hurko, D. Johns et al., MELAS mutation in mtDNA binding site for transcription termination factor causes defects in protein synthesis and in respiration but no change in levels of upstream and downstream mature transcripts. Proc. Natl. Acad. Sci. **89**(10), 4221–4225 (1992)
63. N.J. Gross, G.S. Getz, M. Rabinowitz, Apparent turnover of mitochondrial deoxyribonucleic acid and mitochondrial phospholipids in the tissues of the rat. J. Biol. Chem. **244**(6), 1552–1562 (1969)
64. R. Huemer, K.D. Lee, A. Reeves, C. Bickert, Mitochondrial studies in senescent mice - II. Specific activity, buoyant density, and turnover of mitochondrial DNA. Exp. Gerontol. **6**(5), 327–334 (1971)
65. H. Korr, C. Kurz, T. Seidler, D. Sommer, C. Schmitz, Mitochondrial DNA synthesis studied autoradiographically in various cell types in vivo. Braz. J. Med. Biol. Res. **31**(2), 289–298 (1998)
66. J.P. Burgstaller, I.G. Johnston, N.S. Jones, J. Albrechtova, T. Kolbe, C. Vogl et al., MtDNA segregation in heteroplasmic tissues is common in vivo and modulated by haplotype differences and developmental stage. Cell Rep. **7**(6), 2031–2041 (2014)
67. R.W. Taylor, M.J. Barron, G.M. Borthwick, A. Gospel, P.F. Chinnery, D.C. Samuels et al., Mitochondrial DNA mutations in human colonic crypt stem cells. J. Clin. Invest. **112**(9), 1351–1360 (2003)
68. A. Kowald, M. Dawson, T.B. Kirkwood, Mitochondrial mutations and ageing: can mitochondrial deletion mutants accumulate via a size based replication advantage? J. Theor. Biol. **340**, 111–118 (2014)
69. A.J. Berk, D.A. Clayton, Mechanism of mitochondrial DNA replication in mouse L-cells: asynchronous replication of strands, segregation of circular daughter molecules, aspects of topology and turnover of an initiation sequence. J. Mol. Biol. **86**(4), 801–824 (1974)
70. D.A. Clayton, Replication of animal mitochondrial DNA. Cell **28**(4), 693–705 (1982)
71. A.M. El Zawily, M. Schwarzlaender, I. Finkemeier, I.G. Johnston, A. Benamar, Y. Cao et al., FRIENDLY regulates mitochondrial distribution, fusion, and quality control in Arabidopsis. Plant Physiol. **166**(2), 808–828 (2014)
72. V.M. Sukhorukov, D. Dikov, A.S. Reichert, M. Meyer-Hermann, Emergence of the mitochondrial reticulum from fission and fusion dynamics. PLoS Comput Biol. **8**(10), e1002745 (2012)
73. I.G. Johnston, B. Gaal, R.P. das Neves, T. Enver, F.J. Iborra, N.S. Jones, Mitochondrial variability as a source of extrinsic cellular noise. PLoS Comput. Biol. **8**(3), e1002416 (2012)
74. M. Schwarzländer, D.C. Logan, I.G. Johnston, N.S. Jones, A.J. Meyer, M.D. Fricker et al., Pulsing of membrane potential in individual mitochondria: a stress-induced mechanism to regulate respiratory bioenergetics in Arabidopsis. Plant Cell **24**(3), 1188 (2012)
75. P.F. Chinnery, D.C. Samuels, Relaxed replication of mtDNA: a model with implications for the expression of disease. Am. J. Hum. Genet. **64**(4), 1158–1165 (1999)
76. G.J. Capps, D.C. Samuels, P.F. Chinnery, A model of the nuclear control of mitochondrial DNA replication. J. Theor. Biol. **221**(4), 565–583 (2003)
77. S.E. Durham, D.C. Samuels, L.M. Cree, P.F. Chinnery, Normal levels of wild-type mitochondrial DNA maintain cytochrome c oxidase activity for two pathogenic mitochondrial DNA mutations but not for m. 3243A→G. Am. J. Hum. Genet. **81**(1), 189–195 (2007)
78. I.G. Johnston, N.S. Jones, Evolution of cell-to-cell variability in stochastic, controlled, heteroplasmic mtDNA populations. Am. J. Hum. Genet. **99**(5), 1150–1162 (2016)
79. Mouli PK, Twig G, Shirihai OS. Frequency and selectivity of mitochondrial fusion are key to its quality maintenance function. Biophys. J. **96**(9), 3509–3518 (2009)
80. P.K. Patel, O. Shirihai, K.C. Huang, Optimal dynamics for quality control in spatially distributed mitochondrial networks. PLoS Comput. Biol. **9**(7), e1003108 (2013)
81. M.T. Figge, A.S. Reichert, M. Meyer-Hermann, H.D. Osiewacz, Deceleration of fusion–fission cycles improves mitochondrial quality control during aging. PLoS Comput Biol. **8**(6), e1002576 (2012)

82. Z.Y. Tam, J. Gruber, B. Halliwell, R. Gunawan, Mathematical modeling of the role of mitochondrial fusion and fission in mitochondrial DNA maintenance. PLoS One **8**(10), e76230 (2013)
83. D.T. Gillespie, Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81**(25), 2340–2361 (1977)
84. Z.Y. Tam, J. Gruber, B. Halliwell, R. Gunawan, Context-dependent role of mitochondrial fusion-fission in clonal expansion of mtDNA mutations. PLoS Comput Biol. **11**(5), e1004183 (2015)
85. G. Twig, A. Elorza, A.J. Molina, H. Mohamed, J.D. Wikstrom, G. Walzer et al., Fission and selective fusion govern mitochondrial segregation and elimination by autophagy. EMBO J. **27**(2), 433–446 (2008)
86. H.K. Rajasimha, P.F. Chinnery, D.C. Samuels, Selection against pathogenic mtDNA mutations in a stem cell population leads to the loss of the 3243A→G mutation in blood. Am. J. Hum. Genet. **82**(2), 333–343 (2008)
87. N.G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, vol. 1 (Elsevier, Amsterdam, 1992)
88. R. Grima, P. Thomas, A.V. Straube, How accurate are the nonlinear chemical Fokker–Planck and chemical Langevin equations? J. Chem. Phys. **135**(8), 084103 (2011)
89. J.B. Stewart, P.F. Chinnery, The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat. Rev. Genet. **16**(9), 530–542 (2015)
90. Bratic A, Larsson NG. The role of mitochondria in aging. J. Clin. Invest. **123**(3), 951–957 (2013)
91. I.N. Shokolenko, L.G. Wilson, F.M. Alexeyev, Aging: a mitochondrial DNA perspective, critical analysis and an update. World J. Exp. Med. **4**(4), 46–57 (2014)
92. K. Khrapko, D. Turnbull, Mitochondrial DNA mutations in aging. Prog. Mol. Biol. Transl. Sci. **127**, 29–62 (2014)
93. P.F. Chinnery, D.C. Samuels, J. Elson, D.M. Turnbull, Accumulation of mitochondrial DNA mutations in ageing, cancer, and mitochondrial disease: is there a common mechanism? Lancet **360**(9342), 1323–1325 (2002)
94. D.C. Wallace, D. Chalkia, Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. Cold Spring Harb. Perspect. Biol. **5**(11), a021220 (2013)
95. L.M. Cree, D.C. Samuels, S.C. de Sousa Lopes, H.K. Rajasimha, P. Wonnapinij, J.R. Mann et al., A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. Nat. Genet. **40**(2), 249–254 (2008)
96. H. Cenettlsr, P. McGill, Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA. Nat. Genet. **14**, 146–151 (1996)
97. D. Brown, D. Samuels, E. Michael, D. Turnbull, P. Chinnery, Random genetic drift determines the level of mutant mtDNA in human primary oocytes. Am. J. Hum. Genet. **68**(2), 533–536 (2001)
98. J.N. Wolff, D.J. White, M. Woodhams, H.E. White, N.J. Gemmell, The strength and timing of the mitochondrial bottleneck in salmon suggests a conserved mechanism in vertebrates. PLoS One **6**(5), e20522 (2011)
99. N. Howell, S. Halvorson, I. Kubacka, D. McCullough, L. Bindoff, D. Turnbull, Mitochondrial gene segregation in mammals: is the bottleneck always narrow? Nat. Genet. **90**(1–2), 117–120 (1992)
100. L. Cao, H. Shitara, T. Horii, Y. Nagao, H. Imai, K. Abe et al., The mitochondrial bottleneck occurs without reduction of mtDNA content in female mouse germ cells. Nat. Genet. **39**(3), 386–390 (2007)
101. T. Wai, D. Teoli, E.A. Shoubridge, The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. Nat. Genet. **40**(12), 1484–1488 (2008)
102. I.G. Johnston, J.P. Burgstaller, V. Havlicek, T. Kolbe, T. Rülicke, G. Brem et al., Stochastic modelling, Bayesian inference, and new in vivo measurements elucidate the debated mtDNA bottleneck mechanism. Elife **4**, e07464 (2015)

103. K. Csilléry, M.G. Blum, O.E. Gaggiotti, O. François, Approximate Bayesian computation (ABC) in practice. Trends Ecol. Evol. **25**(7), 410–418 (2010)
104. I.G. Johnston, Efficient parametric inference for stochastic biological systems with measured variability. Stat. Appl. Genet. Mol. Biol. **13**(3), 379–390 (2014)
105. S. Wright, *Evolution and the Genetics of Population. The Theory of Gene Frequencies*, vol. 2 (University of Chicago Press, Chicago, 1987)
106. M. Kimura, Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. U. S. A. **41**(3), 144–150 (1955)
107. I.G. Johnston, N.S. Jones, Closed-form stochastic solutions for non-equilibrium dynamics and inheritance of cellular components over many cell divisions. Proc. R. Soc. A. R. Soc. **471**, 20150050 (2015)
108. J. Poulton, V. Macaulay, D. Marchington, Is the bottleneck cracked? Am. J. Hum. Genet. **62**(4), 752–757 (1998)
109. Solignac M, Génermont J, Monnerot M, Mounolou JC. Genetics of mitochondria in Drosophila: mtDNA inheritance in heteroplasmic strains of D. mauritiana. Mol. Gen. Genet. MGG **197**(2), 183–188 (1984)
110. P. Wonnapinij, P.F. Chinnery, D.C. Samuels , Previous estimates of mitochondrial DNA mutation level variance did not account for sampling error: comparing the mtDNA genetic bottleneck in mice and humans. Am. J. Hum. Genet. **86**(4), 540–550 (2010)
111. C.W. Birky Jr., The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu. Rev. Genet. **35**(1), 125–148 (2001)
112. P. Wonnapinij, P.F. Chinnery, D.C. Samuels, The distribution of mitochondrial DNA heteroplasmy due to random genetic drift. Am. J. Hum. Genet. **83**(5), 582–593 (2008)
113. C. Gardiner, *Stochastic Methods* (Springer, Berlin, 2009)
114. B.L. Gitschlag, C.S. Kirby, D.C. Samuels, R.D. Gangula, S.A. Mallal, M.R. Patel, Homeostatic responses regulate selfish mitochondrial genome dynamics in *C. elegans* (2016). bioRxiv 050930
115. J.P. Burgstaller, I.G. Johnston, J. Poulton, Mitochondrial DNA disease and developmental implications for reproductive strategies. Mol. Hum. Reprod. **21**(1), 11–22 (2015)
116. Third scientific review of the safety and efficacy of methods to avoid mitochondrial disease through assisted conception: 2014 update; June (2014). HFEA
117. Three's company; 9 July 2016. The Economist
118. A. Diot, E. Dombi, T. Lodge, C. Liao, K. Morten, J. Carver et al., Modulating mitochondrial quality in disease transmission: towards enabling mitochondrial DNA disease carriers to have healthy children. Biochem. Soc. Trans. **44**, 1091–1100 (2016)
119. A. Trifunovic, A. Wredenberg, M. Falkenberg, J.N. Spelbrink, A.T. Rovio, C.E. Bruder et al., Premature ageing in mice expressing defective mitochondrial DNA polymerase. Nature **429**(6990), 417–423 (2004)
120. E.C. Røyrvik, J.P. Burgstaller, I.G. Johnston, mtDNA diversity in human populations highlights the merit of haplotype matching in gene therapies. Mol. Hum. Reprod. **22**(11), 809–817 (2016)

Modeling and Stochastic Analysis of the Single Photon Response

Jürgen Reingruber and David Holcman

1 Introduction

Signal transduction at a single molecular level is based on stochastic biochemical events occurring in constrained cellular microdomains. Molecular fluctuations in the transduction pathway generate a cellular background noise, which sets the limit of cell detection. This limit is generic to most of transduction mechanisms that consist of converting a molecular signal into a cellular response. For photoreceptors, light (photons) is transformed into a cellular change of the voltage potential called a hyperpolarization (decrease of the voltage) due to the exit of ions. For olfactory cells, a single odorant molecule can activate a flow of ions through voltage gated channels. During synaptic transmission, neurotransmitters generate a local depolarization. Finally, a transcription factor in the cell nucleus activates or regulates genes, leading to protein expression. In all of these examples a molecular signal leads to a cellular response, but how such a signal overcome the noise and what is the nature of the molecular and cellular noise. We explore these questions based on modeling and analysis of the single photon response in photoreceptors.

A key step in the cellular response to a small molecular event is the amplification of the signal, which occurs by a protein (G-protein) cascades. Of all the G-protein cascades in nature, the best-understood are those initiated by the absorption of a photon in *Drosophila* microvilli [1, 2] and in the outer segment (OS) of vertebrate

J. Reingruber
INSERM U1024; Applied Mathematics and Computational Biology, IBENS, Ecole Normale Supérieure, 46 rue d'Ulm, 75005 Paris, France

D. Holcman (✉)
Institute for Biology École Normale Supérieure, Applied Mathematics and Computational Biology, Paris, France

Churchill College, University of Cambridge, Storey's Way, Cambridge CB3 0DS, UK
e-mail: david.holcman@ens.fr

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_14

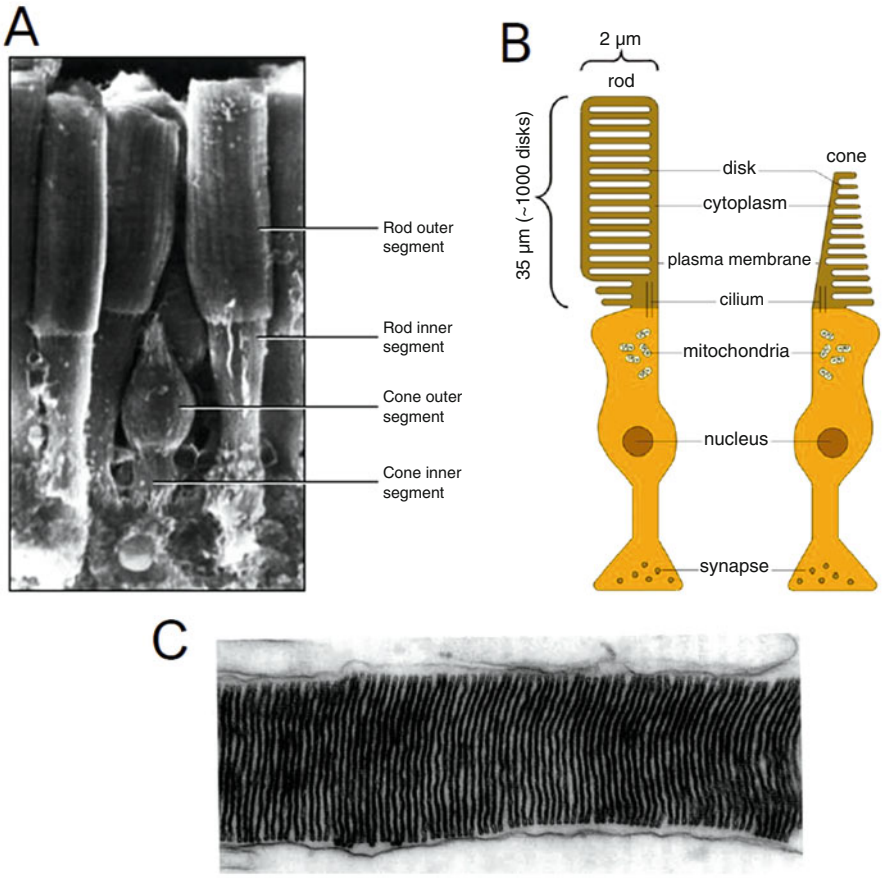


Fig. 1 Geometrical organization of rod and cone photoreceptors. (**a**) Electron microcopy (EM) image of rods and cones located in the retina. (**b**) Schematic modeling of a rod and a cone showing their polarized structure: light sensitive outer segment, inner segment with nucleus and synaptic terminal. (**c**) Cross section of a rod outer segment: internal disks divide the outer segment into almost independent cylindrical compartments

rod photoreceptors (Fig. 1) [1, 3, 4]. The rods of amphibians and mammals have been shown to have the remarkable ability to detect single photons of light above background noise [5, 6]. But amphibian and mammalian rods differ in concentrations and biochemical properties of proteins involved in the light response, and by as much as an order of magnitude in the diameter of their disk membranes, where the reactions of the cascade take place. It remains largely unknown how the biochemistry and the geometry adapt to guarantee a reliable macroscopic response initiated by a single molecular event.

We summarize in this review recent progress in mathematical modeling of single photon response in rod photoreceptors. The modeling, analysis, and simulations combine several methods. First, because it is not yet possible to model millions of interacting molecules, the three-dimensional geometry of the rod geometry is

reduced to a one dimension. This is possible because diffusion in a thin cylinder is well approximated by a one dimension process. In that context, reaction–diffusion equations can be written for the subcellular molecular interactions occurring inside the rod structure. Second, there is a geometrical separation between chemical reactions occurring on the membrane and others inside the three-dimensional cytoplasm. This geometrical separation allows studying separately two- and three-dimensional chemical reactions. The two-dimensional chemical reactions do not suffer from geometrical confinement and are can be studied using Markov chains. However, connecting the output of two-dimensional reactions with the three-dimensional ones uses the one-dimensional diffusion reduction approximation. The overall reduced modeling allows to perform stochastic simulations that explain the variability in the biochemistry and allows to study the major sources of noise during a single photon response.

We recall briefly that noise in the photoreceptor is generated by the fluctuations in the activity of a critical enzyme called phosphodiesterase (PDE). This enzyme fulfills two essential functions. First, the phosphodiesterase that becomes activated through the transduction cascade after a photon absorption (light-activated PDE) increases the hydrolysis of cGMP, a diffusible second messenger controlling the opening of ionic membrane channels, leading to channel closure and cell hyperpolarization; second, spontaneously activated PDE is necessary to maintain in darkness a steady-state cGMP concentration and to set the cGMP turnover rate, an important determinant of the time scale of the photon response [7, 8] (Fig. 2). Fluctuations in the number of spontaneously activated PDEs generate a background noise that is commonly referred to as the dark noise [7, 9, 10]. The main source of variability in the amplitude of the single photon response is due to variability in the number of light-activated PDEs [6, 11–13].

Photon response curve and noise generated within the transduction cascade are evaluated using spatially resolved reaction–diffusion equations and stochastic simulations of PDE activations at the level of single molecules. We present here a summary of multiscale simulations that account for the molecular details (PDE activations and cGMP hydrolysis) and the intrinsic molecular noise called dark noise. The result of the simulations can be directly compared to experimental recordings and the analytical expressions for the dark noise power spectrum are used to extract the values of key parameters from the analysis of measured current recorded not only in wild type (WT) but also in genetically modified cells such as $\text{Caps}^{-/-}$ knockout mice.

This review is organized as follows: in the first part, we present the homogenization procedure to reduce the three-dimensional rod outer-segment geometry to a one dimensional with an effective diffusion coefficient. In the second part, Markov chains are used for modeling the stochastic activation of PDE molecules following a photon absorption. We also present the modeling of the spontaneous PDE activation. In the third, we analyze the Laplace equation for computing the cGMP hydrolysis rate, based on the narrow escape theory in narrow band [14]. In Sect. 4, we introduce the coupled system of equations for cGMP and calcium currents. In Sect. 5, we present the stochastic simulations of a single photon response. Finally, in Sect. 7, we explain how numerical simulations are used to extract biophysical parameters from dark noise recordings and single photon response.

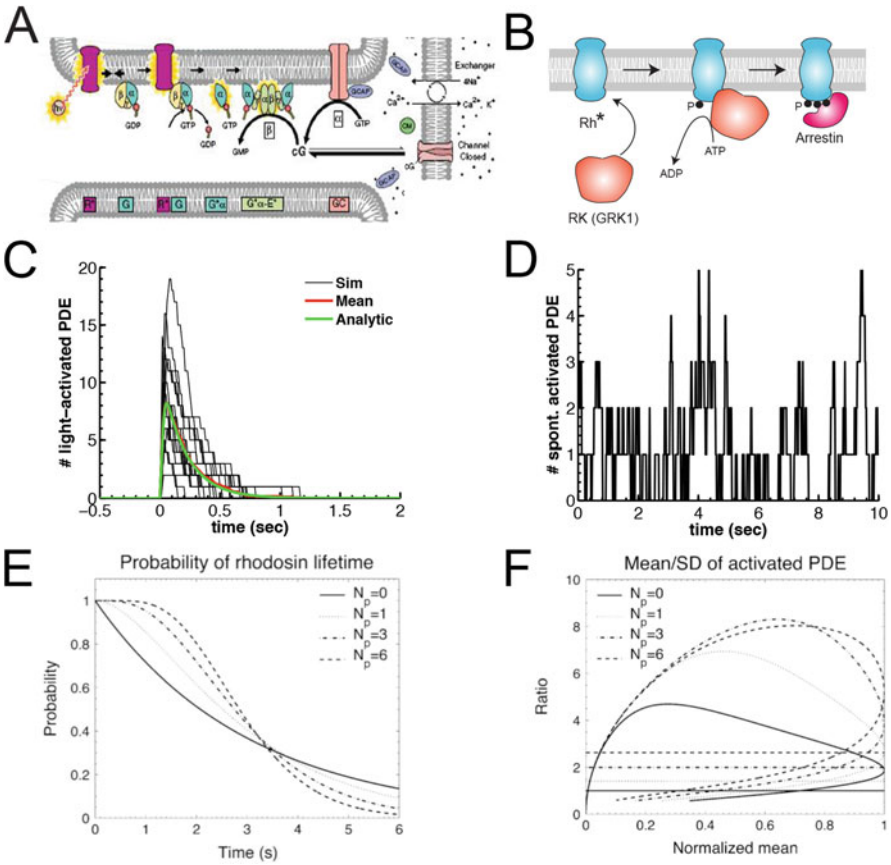


Fig. 2 Signal transduction and PDE activation. (**a**) Signal transduction cascade of a vertebrate photoreceptor. (**b**) Schematic representation of deactivation of an activated rhodopsin following multiple phosphorylations and arrestin binding [18]. (**c**) Stochastic simulations (*black*) of the number of activated PDEs after a single photon absorption in a mouse rod, average (*red*) and the analytic result for the mean (*green*). (**d**) Stochastic simulation of the number of spontaneously activated PDEs in a mouse compartment with a mean $\bar{P}^*_{\text{sp,comp}} = 0.9$ and $\mu_{\text{sp}} = 12.4s^{-1}$. (**e**) Probability of activated rhodopsin lifetime depending on the number of phosphorylation sites. (**f**) Mean to standard-deviation ratio for activated PDEs plotted as a function of the mean PDE normalized by its maximum value [15]

2 Modeling Phosphodiesterase (PDE) Activation After a Photon Absorption Using Markov Chain

To simulate the time course of the stochastic number of activated PDE P^* following a single photon absorption, we use Markov chain [15] (see also [16]): after photon absorption, a rhodopsin molecule undergoes a conformational modification and changes from an inactive R into an active R^*_N state, where N is the total number of

Table 1 Parameters for PDE activation

Parameter	Definition
$\bar{P}^*_{\text{sp,comp}}$	Mean number of spontaneous activated PDE molecules per compartment
$\bar{P}^*_{\text{li,max}}$	Mean of the peak number of light-activated PDE
ρ_{pde}	PDE surface density
ν_{sp}	Spontaneous PDE activation rate
μ_{sp}	Spontaneous PDE deactivation rate
μ_{li}	Deactivation rate for light-activated PDE
τ_{Rh}	Activated rhodopsin lifetime
N_p	Number of rhodopsin phosphorylation steps
$\gamma_{\text{rt,max}}$	Maximal transducin activation rate
ω	Decay rate of transducin activation with the number of phosphorylation steps
γ_{tp}	Rate by which activated transducin activates PDE

available phosphorylation sites. The R^*_N phosphorylation is catalyzed by a rhodopsin kinase (RK) that gradually reduces the activity of rhodopsin (Fig. 2a, b). Through phosphorylation, rhodopsin in the state R^*_n undergoes a transition to state R^*_{n-1}, modeled by the state dependent Poissonian phosphorylation rate λ_n. R^*_n activates the G-protein transducin T^* with rate k_n, which constitutes an amplification process. A T^* transducin binds to a single PDE with a rate μ_t and forms a complex denoted by P^* (see Fig. 2a–c), which subsequently deactivates with rate μ_p. Eventually, the rhodopsin R^*_n becomes deactivated through another molecule arrestin binding with a rate μ_n. The kinetic reactions are summarized as follows (see Table 1):

$$\begin{aligned} R^*_n &\xrightarrow{\lambda_n} R^*_{n-1} \\ R^*_n &\xrightarrow{\mu_n} R \\ R^*_n + T &\xrightarrow{k_n} R^*_n + T^* \\ T^* + P &\xrightarrow{\mu_t} P^* \\ P^* &\xrightarrow{\mu_p} P \end{aligned} \tag{1}$$

The state of the signalling process is described by three stochastic variables integer values (n, l, k), which are the phosphorylation state $0 \le n \le N$ of R^* (n corresponds to the number of remaining unphosphorylated sites), the number $0 \le l \le \infty$ of $T*$ and $0 \le k \le \infty$ for $P*$. The joint probability $P_n(l, k, t)$ satisfies the Master equation

$$\begin{aligned} \frac{\partial}{\partial t} P_n(l,k,t) = {} & \lambda_{n+1} P_{n+1}(l,k,t) + k_n P_n(l-1,k,t) \\ & + \mu_t (l+1) P_n(l+1,k-1,t) + \mu_p (k+1) P_n(l,k+1,t) \\ & - \left(\lambda_n + \mu_n + k_n + \mu_t l + \mu_p k\right) P_n(l,k,t). \end{aligned} \tag{2}$$

In state $n = 0$, all sites are phosphorylated and $\lambda_0 = 0$. After a photon absorption, R^* is in state $n = N$ and the number of T^* and P^* is zero. The initial condition is given by $P_n(l, k, 0) = \delta_{n,N}\delta_{l,0}\delta_{k,0}$, where ($\delta_{i,j}$ is the Kronecker symbol).

2.1 First and Second Moment (Mean and Variance) of Rhodopsin Lifetime Distribution

The mean and variance of R^* lifetime can be computed from the probabilities $P_n(t)$ to find $R*$ in state n at time t. By summing Eq. (2) over l and k, the probability vector $\vec{P}(t) = (P_N(t), \ldots, P_0(t))^\top$ satisfies the equation

$$\frac{d}{dt}\vec{P}(t) = \mathbf{S}\,\vec{P}(t) \quad \text{with} \quad \mathbf{S} = \begin{pmatrix} -\beta_N & & & \\ \lambda_N & -\beta_{N-1} & & \\ & & \vdots & \\ & & \lambda_1 & -\beta_0 \end{pmatrix} \tag{3}$$

and $\beta_n = \lambda_n + \mu_n$. To compute the mean R^* lifetime, we integrate Eq. (3) using the initial condition $\vec{P}(0) = (1, \ldots, 0)^\top$ and we use that $\vec{P}(t)$ vanishes for $t \to \infty$. We obtain for the mean time

$$\bar{\tau} = \sum_{n=0}^{N} \int_0^\infty P_n(t)dt = -\mathrm{Tr}\left(\mathbf{S}^{-1}\vec{P}(0)\right) = \sum_{n=0}^{N} \frac{1}{\beta_n} \prod_{k=n+1}^{N} p_k\,, \tag{4}$$

with $p_n = \frac{\lambda_n}{\beta_n}$. Equation (4) has an intuitive interpretation: it is the sum of mean lifetimes $\frac{1}{\beta_n}$ in each state n multiplied by the probability to reach this state before being deactivated via arrestin binding (see also Fig. 2e).

The variance is computed by integration by parts in the relation

$$\begin{aligned} \Sigma_\tau &= -\sum_{n=0}^{N} \int_0^\infty t^2 \frac{d}{dt} P_n(t)dt - \bar{\tau}^2 = \sum_{n=0}^{N} \int_0^\infty 2tP_n(t)dt - \bar{\tau}^2 = 2\mathrm{Tr}\left(\mathbf{S}^{-2}\vec{P}(0)\right) - \bar{\tau}^2 \\ &= 2\sum_{n=0}^{N}\sum_{j=0}^{n} \frac{1}{\beta_n}\frac{1}{\beta_j} \prod_{k=j+1}^{N} p_k - \bar{\tau}^2\,. \end{aligned} \tag{5}$$

The coefficient of variation (CV) of $R*$ lifetime (Fano factor) has a lower bound that depends only on the number of phosphorylation sites N [15]. Indeed using Eqs. (4) and (5), we have

$$CV_\tau = \frac{\sqrt{\Sigma_\tau}}{\bar{\tau}} \geq \frac{1}{\sqrt{N}}\,. \tag{6}$$

The minimum $CV_\tau = \frac{1}{\sqrt{N}}$ is achieved for $\beta_n = \text{const}$ and $p_n = 1$. The first condition reduces the lifetime variability of the various deactivation states. The latter condition requires that arrestin binds only when R^* is fully phosphorylated, which maximizes the effective number of deactivation steps.

2.2 Stochastic Analysis of the Number of Activated PDE

As shown in the previous paragraph, multiple phosphorylations of the rhodopsin molecule reduce the CV_τ and lead to a more reliable R^* deactivation process. Is a reliable R^* deactivation process leads to a minimal variance in the number of P^*? How multiple phosphorylations affect the mean and variance of the number of activated PDE? In particular, does a low CV_τ entail a low CV of the number of activated PDE? The answer to these questions is based on a system of differential equations to compute numerically the time dependent mean and variance of P^* and the mean and variance of the total number of P^* that are activated during a single photon response. We now present such equations.

2.2.1 System of Differential Equations for Mean and Variance

The mean and variance that depend only on the phosphorylation state n of R^* can be computed by decomposing the matrix $\mathbf{S}$ into a sum of left eigenvectors. By decomposing the activation rate vector $\vec{k} = (k_N, \ldots, k_0)^\top$ into N eigenvectors $\vec{k}_i$ of the matrix $\mathbf{S}$, we get

$$\vec{k} = \sum_{i=0}^{N} \vec{k}_i \quad \text{with} \quad \vec{k}_i^\top \mathbf{S} = -\beta_i \vec{k}_i^\top \,, \tag{7}$$

we obtain for the individual mean values the relation

$$\frac{d}{dt}\bar{k}_i(t) = \frac{d}{dt}\sum_{n=0}^{N} k_{i,n} P_n(t) = \sum_{n,m=0}^{N} k_{i,n}\mathbf{S}_{n,m} P_m(t) = -\beta_i \bar{k}_i(t)\,. \tag{8}$$

Together with the initial condition $\vec{P}(0) = (1, \ldots, 0)^\top$, we get

$$\bar{k}(t) = \sum_{i=0}^{N} \bar{k}_i(t) = \sum_{i=0}^{N} k_{i,N}\, e^{-\beta_i t}\,. \tag{9}$$

Similarly, the variance $\Sigma_k(t) = \sum_{n=0}^{N} k_n^2 P_n(t) - \bar{k}(t)^2$ is calculated by decomposing the vector $\vec{x} = (k_N^2, \ldots, k_0^2)$.

We now present the time dependent mean and variance of PDE defined by

$$\bar{P}(t) = \sum_{n=0}^{N} \sum_{l,k=0}^{\infty} k P_n(l,k,t) \quad \text{and} \quad \Sigma_p(t) = \sum_{n=0}^{N} \sum_{l,k=0}^{\infty} k^2 P_n(l,k,t) - \bar{P}(t)^2. \quad (10)$$

Using Eq. (2), it is possible to obtain a closed system of differential equations for the mean and cross-correlations,

$$\begin{aligned}
\frac{d}{dt}\bar{T}(t) &= -\mu_t \bar{T}(t) + \bar{k}(t) \\
\frac{d}{dt}\bar{P}(t) &= -\mu_p \bar{P}(t) + \mu_t \bar{T}(t) \\
\frac{d}{dt}\Sigma_t(t) &= -2\mu_t \Sigma_t(t) + \mu_t \bar{T}(t) + 2\Sigma_{kt}(t) + \bar{k}(t) \\
\frac{d}{dt}\Sigma_p(t) &= -2\mu_p \Sigma_p(t) + 2\mu_t \Sigma_{\mathrm{tp}}(t) + \mu_t \bar{T}(t) + \mu_p \bar{P}(t) \\
\frac{d}{dt}\Sigma_{\mathrm{tp}}(t) &= -(\mu_t + \mu_p)\Sigma_{\mathrm{tp}}(t) + \mu_t \Sigma_t(t) - \mu_t \bar{T}(t) + \Sigma_{kp}(t)\,.
\end{aligned} \quad (11)$$

The mean activation rate $\bar{k}(t)$ can be computed independently and therefore is an input function.

To close this system we need additional equations for $\Sigma_{kp}(t)$ and $\Sigma_{kt}(t)$. Using the decomposition in Eq. (7), we write

$$\Sigma_{kt}(t) = \sum_i \Sigma_{k_i t}(t) \quad \text{and} \quad \Sigma_{kp}(t) = \sum_i \Sigma_{k_i p}(t)\,. \quad (12)$$

Finally, the missing equations that close the system are

$$\begin{aligned}
\frac{d}{dt}\Sigma_{k_i t}(t) &= -(\beta_i + \mu_t)\Sigma_{k_i t}(t) + \sum_j \Sigma_{k_i k_j}(t) \\
\frac{d}{dt}\Sigma_{k_i p}(t) &= -(\beta_i + \mu_p)\Sigma^2_{k_i p}(t) + \mu_t \Sigma_{k_i t}(t)\,.
\end{aligned} \quad (13)$$

The correlation functions $\Sigma_{k_i k_j}(t) = \sum_n k_{i,n} k_{j,n} P_n(t) - \bar{k}_i(t)\bar{k}_j(t)$ and $\bar{k}(t)$ are known functions.

2.3 *Stochastic Dynamics of the Number of Activated PDE*

To further investigate how the variability of $R*$ deactivation can influence the production of P^*, it is useful to compute the mean $\bar{P}_{\mathrm{tot}}$ and variance $\Sigma_{p_{\mathrm{tot}}}$ of the total number of P^* produced during a SPR. This computation is obtained by setting

the P^* deactivation rate to zero, $\mu_p = 0$, in which case all P^* are conserved. We obtain

$$\bar{P}_{\text{tot}} = \int_0^\infty \bar{k}(t)dt = \sum_{n=0}^{N} \frac{k_n}{\beta_n} \prod_{k=n+1}^{N} p_k \,, \tag{14}$$

$$\Sigma_{p_{\text{tot}}} = \bar{P}_{\text{tot}} + 2\int_0^\infty \Sigma_{kt}(t)dt = \bar{P}_{\text{tot}} + \sum_{n=0}^{N}\sum_{j=0}^{n} \frac{k_n}{\beta_n}\frac{k_j}{\beta_j} \prod_{k=j+1}^{N} p_k - \bar{P}_{\text{tot}}^2 \,. \tag{15}$$

The lower bound for the CV of the total number of P^* is

$$\text{CV}_{p_{\text{tot}}} = \frac{\sqrt{\Sigma_{p_{\text{tot}}}}}{\bar{P}_{\text{tot}}} \geq \frac{\sqrt{1 + \frac{N}{\bar{P}_{\text{tot}}}}}{\sqrt{N}} \,. \tag{16}$$

Although the coefficient of variations $\text{CV}_{p_{\text{tot}}}$ and CV_τ share the same lower bound $1/\sqrt{N}$, in general, minimal values for both cannot be achieved simultaneously. Indeed, whereas a minimal $\text{CV}_{p_{\text{tot}}}$ requires $k_n/\beta_n = \text{const}$ and $p_n = 1$, a minimal CV_τ is achieved for $1/\beta_n = \text{const}$ and $p_n = 1$. Thus, by adjusting the activation rates k_n one can have a minimal $\text{CV}_{p_{\text{tot}}}$ even when CV_τ is far from being minimal. Thus, a reliable $R*$ lifetime is neither necessary nor sufficient for a reliable PDE activation. For constant transducin activation rates ($k_n = k$) we have almost linear relations $\bar{P}_{\text{tot}} = k\bar{\tau}$, $\Sigma_{p_{\text{tot}}}$ and $\Sigma_{p_{\text{tot}}} = \bar{P}_{\text{tot}} + k^2\Sigma_\tau$ (see also [15]).

2.4 Modeling the Spontaneous PDE Activation

In addition to PDE activation after a photon absorption, PDE molecules activate and deactivate spontaneously with Poisson rates ν_{sp} and μ_{sp} according to the biochemical reaction

$$P \underset{\mu_{\text{sp}}}{\overset{\nu_{\text{sp}}}{\rightleftharpoons}} P^*_{\text{sp}} \,. \tag{17}$$

Spontaneous PDE activation is an inconvenient mechanism, generating a background noise (dark noise) that obscures the signal from a single photon absorption: the signal generated by a photon has to overcome the dark noise amplitude [7]. However, spontaneous PDE activations are crucial because they hydrolyze cGMP in the dark and are essential to guarantee a steady-state concentration of cGMP in the transduction current in the dark.

Using Eq. (17), the average steady-state number of spontaneously activated PDE in a compartment is given by

$$\bar{P}^*_{\text{sp,comp}} = 2\rho_{\text{pde}}\pi R^2 \frac{\nu_{\text{sp}}}{\mu_{\text{sp}}} \,, \tag{18}$$

where ρ_{pde} is the PDE surface density and R is the compartment radius. For example, for toad rods, assuming $\nu_{\text{sp}} = 4 \times 10^{-4}\,\text{s}^{-1}$ and $\mu_{\text{sp}} = 1.8\,\text{s}^{-1}$ [7], $R = 3\,\mu\text{m}$ and $\rho_{\text{pde}} = 100\,\mu\text{m}^{-2}$ (see Table 3), we find $\bar{P}^*_{\text{sp,comp}} = 1.25$. Figure 2d shows a simulation of the stochastic number of spontaneously activated PDE in a mouse compartment with $\bar{P}^*_{\text{sp,comp}} = 0.9$ and $\mu_{\text{sp}} = 12.4\,\text{s}^{-1}$ (see Table 3).

We use Eq. (17) together with the SSA Gillespie algorithm [17] to simulate the time course of the stochastic number of spontaneously activated PDE in a compartment (Fig. 2d).

2.5 *Homogenization of the Three-Dimensional ROS Geometry and Reduction of Diffusion in a Long Cylinder*

The outer segment (OS) is the sensory unit of the photoreceptor. The rod OS is divided by internal parallel disks into compartments connected to each other through narrow gaps between the disk rim and the OS membrane and through incisures (Fig. 3). A photon is absorbed by a rhodopsin photopigment attached to the surface of a single internal disk. As discussed in the previous paragraphs, rhodopsin activation after a photon absorption triggers the activation of many PDE enzymes via a G-protein (transducin) coupled amplification cascade. Because PDE and transducin molecules are also attached to the disk surface, the activation process occurs on the internal disk surface, where the photon has been absorbed (Fig. 2a). An activated PDE hydrolyzes (kill) the cytosolic second messenger cGMP that controls the opening of CNG ion channels in the OS membrane.

The compartmentalization of the OS restricts the diffusion of cGMP between neighboring compartments, whereas diffusion within a compartment is not affected and leads to rapid transversal equilibration. We therefore adopt the approximation of a transversally well stirred OS where the three-dimensional geometry is reduced to an effective one-dimension model with an effective longitudinal diffusion constant [19, 20]. We now describe this geometrical reduction as shown in Fig. 3 and the estimation of the cGMP hydrolysis rate using a general Narrow Escape Theory [21] in a flat cylinder that involves two- and three-dimensional asymptotic estimates [14].

2.6 *Computing the Effective Longitudinal Diffusion Constant*

The model of diffusion reduction starts with considering Brownian particles in the OS, that are driven by thermal noise and a trajectory in the cytoplasmic fluid is well described by the overdamped approximation (Smoluchowski limit) of the Langevin equation. For a molecule located at position $X(t)$, the velocity satisfies the stochastic equation

$$\gamma \dot{X} + F(X) = \sqrt{2\gamma\varepsilon}\dot{\mathbf{w}} \tag{19}$$

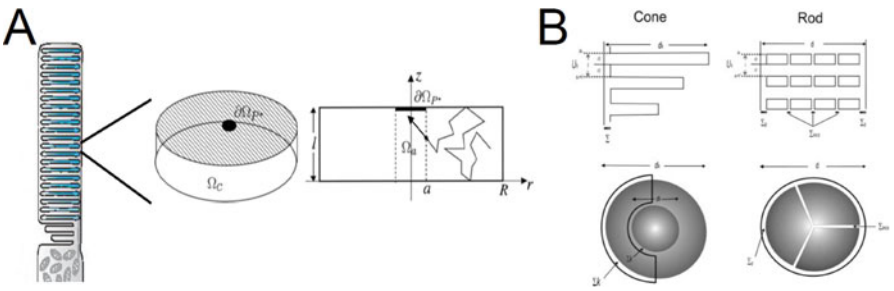


Fig. 3 PDE hydrolysis rate and homogenization. (**a**) Schematic representation of a cylindrical compartment with an activated PDE on the surface. A diffusing cGMP trajectory in the cytosol is terminated (hydrolyzed) when reaching the activated PDE site for the first time. (**b**) Schematic representation of cone and rod outer segment structure used to compute an effective longitudinal diffusion constant

where $F(x)$ are forces applied onto the particle, γ is the viscosity coefficient, $\varepsilon = \frac{kT}{m}$ is the thermal noise, and $\dot{w}$ is the white noise produced by thermal collisions [20, 22].

To study Eq. (19), we make three assumptions: (1) particles do not bind; (2) in the time scale of seconds, short range electrostatic interactions that arise from the charged disc membrane surfaces [23] and/or charged particles in solution average and cancel out. As a consequence all electrostatic terms are neglected and the force term $F(x)$ in Eq. (19) is set to zero; (3) particles do not permeate across OS membranes.

From the general theory of diffusion, it is well known that the probability density function (pdf) of one molecule associated with Eq. (1) satisfies the standard three-dimensional diffusion equation inside the cytoplasmic fluid. Under the assumption of independent molecules, the concentration is simply the product of the pdf by the number of molecule and satisfies the diffusion equation within the OS:

$$\frac{\partial c}{\partial t} = D\Delta c \tag{20}$$

$$c(x, 0) = c_0(x) \tag{21}$$

where $c_0(x)$ is some initial concentration, and D is the diffusion constant.

2.7 *Longitudinal Diffusion in Rod Outer Segments*

ROS consists of repeating spatial compartments, U_k (Fig. 3b). Each compartment comprises the distance from one disc surface to the comparable surface in the next disc. The repeat distance is l, and it consists in two parts: the cytoplasmic space separating two adjacent discs (interdisc space, dimension $= l/2$) and the disc itself (dimension $= l/2$). Diffusion between adjacent interdisc spaces occurs through

either disc incisures or the perimeter gap that separates disc edges from the plasma membrane. The compartments' radius is constant and denoted by r. We adopt the following notation: N_k is the number of free Brownian particles in the U_k compartment of volume V_k. The present analysis originates from [20].

The objective of the following derivation is to compute the ROS longitudinal diffusion constant, D_l, in terms of the OS's structure. Variation in the number of particles in unit U_k equals the difference of flux into compartments $k-1$ and $k+1$, that is

$$\frac{dN_k(t)}{dt} = -D[J_k - J_{k+1}] \tag{22}$$

By definition, J_{k+1} is the flux between unit U_k and U_{k+1} through Σ_{k+1}, the open surface that joins them,

$$J_{k+1} = \int_{\Sigma_{k+1}} \frac{\partial c(x,t)}{\boldsymbol{n}(x)} dS \tag{23}$$

where $\boldsymbol{n}(x)$ is the normal derivative pointing outside U_k and $\partial U_k = \Sigma_{k+1} \cup \Sigma_k$. In the time scale of seconds, the concentration is assumed to be uniform within each U_k compartment. This assumption is valid because: (1) diffusion within a compartment (transverse diffusion) is the standard two-dimensional diffusion process where the diffusion constant equals to the aqueous diffusion constant; (2) the ratio of the absorbing boundary surface divided by the reflective boundary surface of a given compartment is small.

As a consequence, the diffusion along the longitudinal axis is much slower than along the transverse axis. Finally since the concentration inside a compartment equilibrates quickly at the time scale of seconds, the concentration can be considered to be uniform. Thus the flux through disk incisures and the perimeter gap is the same and does not depend on the transverse spatial variable.

Combining Eqs. (22) and (23) yields,

$$\frac{1}{V_k}\frac{dN_k(t)}{dt} = \frac{D}{V_k}(n\Sigma_{\text{incs}} + \Sigma_g)\left(\frac{\partial c(x_{k+1},t)}{\partial x} - \frac{\partial c(x_k,t)}{\partial x}\right), \tag{24}$$

where Σ_{incs} is the surface area of a single disk incisures, n is the number of incisures, and Σ_g is the surface of the perimeter gap between the disk's edge and plasma membrane. $\Sigma_g = 2\pi r g_w$, where r is the ROS radius and g_w is the size of the perimeter gap. The concentration at points $x_{k+1} = x_k + l$ and x_k is evaluated by a Taylor expansion. At the first order, since $c(x_k,t) = \frac{N_k(t)}{V_k}$, then

$$\frac{\partial c(x_k,t)}{\partial t} = \frac{1}{V_k}\frac{dN_k(t)}{dt} = \frac{D}{V_k}(n\Sigma_s + \Sigma_l)l\frac{\partial^2 c(x_k,t)}{\partial x_k^2}. \tag{25}$$

The translation invariance of the rod outer-segment geometry implies that the volume V_k of the compartment U_k is constant with respect to k and equals to the free interspace volume (free volume of the unit plus the free volume of the incisures)

$$V_k = \pi r_k^2 l/2 + nV_{\text{incs}} + V_g = \pi r^2 l/2 + (n\Sigma_s + 2\pi r g_w) l/2, \tag{26}$$

where V_{incs} is the volume of the incisure and V_g is the volume of the perimeter gap. Finally, Eq. (25), can be reduced to the form of the standard one-dimensional diffusion equation, for $x \in [0, L]$

$$\frac{\partial c(x,t)}{\partial t} = D_l \frac{\partial^2 c(x,t)}{\partial x^2}, \tag{27}$$

where the longitudinal diffusion constant is defined to be:

$$D_l = \frac{D(n\Sigma_{\text{incs}} + \Sigma_g) l}{\pi r^2 l/2 + (n\Sigma_{\text{incs}} + \Sigma_g) l/2} = 2D \frac{1}{\frac{\pi r^2}{n\Sigma_{\text{incs}} + \Sigma_g} + 1} \tag{28}$$

2.8 Longitudinal Diffusion in Cone Outer Segments (COS)

COS consists of repeating U_k compartments, each comprising the distance from the intracellular surface of one membrane fold to the intracellular surface of the next one (Fig. 3b). The repeat distance is l and it consists of two segments: the membrane fold (size = $l/2$) and the distance separating one fold from the next (size = $l/2$). The volume connecting adjacent folds can be very complicated, however in average, we can assume that the geometrical shape is fixed and well approximated by a cylinder of length = $l/2$ and diameter δ.

To remember that δ is the diameter of a disk of area equal to the area of the real surface, we will refer to δ as the "equivalent diameter." The diameter of a fold at position x_k is denoted by d_k and it increases linearly with the longitudinal coordinate as given by $d_{k+1} = d_k + d_0$, where $d_0$is the incremental distance.

Derivation of the longitudinal diffusion equation in COS proceeds through the same steps as the derivation for ROS, but differs due to the difference in the geometry. This difference is due to variation of the spatial compartment U_k. In COS, the time variation in the particle number in the U_k compartment is given by Eq. (22) and the flux through the Σ_{k+1} surface is given by Eq. (23). From these equations, the variation in time of the concentration in a COS compartment is given by:

$$\frac{1}{V_k} \frac{dN_k(t)}{dt} = \frac{1}{V_k} \left(D\Sigma_{k+1} \frac{\partial c(x_{k+1}, t)}{\partial n(x)} - D\Sigma_k \frac{\partial c(x_k, t)}{\partial n(x)} \right) \tag{29}$$

Since $\Sigma_{k+1} = \Sigma_k$, using a Taylor expansion of the concentration c(x,t), we have

$$\frac{\partial c(x_k,t)}{\partial t} = \frac{1}{V_k}\frac{dN_k(t)}{dt} = \frac{D\Sigma_{k+1}l}{V_k}\frac{\partial^2 c(x_k,t)}{\partial^2 x_k} \tag{30}$$

The area of the surface Σ_{k+1} is given by $\Sigma_{k+1} = \pi(\delta/2)^2$ and the volume is

$$V_k = \pi(\delta/2)^2 l/2 + \pi(d_k/2)^2 l/2 \tag{31}$$

For $d_{\min}$ the smallest COS diameter (at its tip), $d_{\max}$ is the maximal COS diameter (at its base), L is the COS length

$$\alpha = \frac{d_{\max} - d_{\min}}{L} \tag{32}$$

Because of the cone geometry, at position x_k the cone diameter is $d_k = \alpha x_k + d_{\min}$, the longitudinal diffusion equation can now be expressed as:

$$\frac{\partial c(x_k,t)}{\partial t} = \frac{D\pi(\delta/2)^2 l}{\pi(\delta/2)^2 l/2 + l/8\pi(\alpha x_k + d_{\min})^2}\frac{\partial^2 c(x_k,t)}{\partial x_k^2} \tag{33}$$

for $x \in [0, L]$, the equation simplifies

$$\frac{\partial c(x,t)}{\partial t} = \frac{2D\delta^2}{\delta^2 + (x\alpha + d_{\min})^2}\frac{\partial^2 c(x,t)}{\partial x^2} \tag{34}$$

The longitudinal diffusion coefficient (which is now in one dimension a function of x) is explicitly given by

$$D(x) = \frac{2D\delta^2}{\delta^2 + (x\alpha + d_{\min})^2}. \tag{35}$$

2.9 Determination of the Diameter Function $\delta(x)$ from the COS Structure

Using the COS structure, Eq. (33) can be modified to include a spatial dependency in the δ variable. Recall that the COS geometry is characterized by the following global parameters. L is the Length, r_{base} is the radius at the base, r_{tip} is the radius at the tip, and $\alpha' = \frac{r_{\text{base}} - r_{\text{tip}}}{L}$. The surface $\Sigma_g(x)$ at the longitudinal position x, connecting adjacent folds is not circular, rather it is a semicircular disk that surrounds half the

perimeter of the membrane folds over the entire COS length (Fig. 3b). Although by analyzing the electron Microscopy picture, $\Sigma_g(x)$ fluctuates along the OS, we will neglect such fluctuations compared to the mean.

If $r_m(x)$ is the radius up to the plasma membrane and $r_f(x)$ is the radius of the membrane fold at position x, the surface $\Sigma_g(x)$ (Fig. 3b) then the area of $\Sigma_g(x)$ is

$$\Sigma_g(x) = \frac{\pi}{2}(r_m(x)^2 - r_f(x)^2) \tag{36}$$

For $r_{\text{tip}} \leq r_f(x) \leq r_{\text{base}}$, $r_f(x) = r_{\text{tip}} + \alpha' x$ and $r_m(x) = r_f(x) + d$, the gap between the closed loop of a fold and the plasma membrane is about 100 Δ, that is $d = r_m - r_f$ =0.01 μ is small compared to r_m. We can approximate $\Sigma_g(x)$ by

$$\Sigma_g(x) = \pi d r_f(x) = \pi d \left(r_{\text{tip}} + \left(\frac{r_{\text{base}} - r_{\text{tip}}}{L} \right) x \right) \tag{37}$$

Let us define the diameter function $\delta(x)$ of a disk along the COS of the same area as $\Sigma_g(x)$ by

$$\Sigma_g(x) = \pi(\delta(x)/2)^2 \tag{38}$$

then

$$\delta(x) = 2\sqrt{d(r_{\text{tip}} + \alpha' x)} \tag{39}$$

Using the result of the previous section, we can now incorporate the small changes in the diameter function $\delta(x)$ into the cone diffusion equation and Eq. (33) becomes

$$\frac{\partial c(x,t)}{\partial t} = \frac{2D\delta(x)^2}{\delta(x)^2 + (x\alpha + d_{\min})^2} \frac{\partial^2 c(x,t)}{\partial x^2} \tag{40}$$

with

$$\delta(x) = 2\sqrt{d(r_{\text{tip}} + \alpha x)} \tag{41}$$

An other derivation of this result using homogenization method can be found in [24].

3 Computing cGMP Hydrolysis Rates

Key parameters that control the CNG channel opening and the current response are the rates of cGMP hydrolysis by spontaneously and light-activated PDE, denoted by k_{sp} and k_{li} that we shall compute now.

3.1 Rate of Hydrolysis by Spontaneously Activated PDE

The rate of cGMP hydrolysis by a spontaneously activated PDE $k_{\rm sp}$ is computed from formula [19, 25]

$$\beta_d = k_{\rm sp}\bar{P}^*_{\rm sp,comp}\,. \tag{42}$$

For example, for a toad rod with the experimental value $\beta_d = 1\,{\rm s}^{-1}$ [11, 16] and $\bar{P}^*_{\rm sp,comp} = 1.25$ [7] we find $k_{\rm sp} \approx 0.8\,{\rm s}^{-1}$. For a mouse rod with $\beta_d = 4.1\,{\rm s}^{-1}$ [26] and $\bar{P}^*_{\rm sp,comp} = 0.9$ (see further down) we get $k_{\rm sp} = 4.5\,{\rm s}^{-1}$ (Fig. 3). The different values of $k_{\rm sp}$ in toad and mouse might be due to the temperature, due to the different encounter rates between PDE and cGMP, or due to differences in the PDE enzyme between amphibians and mammals.

3.2 Computing the Rate of Hydrolysis by Light-Activated PDE from the Narrow Escape Theory

Light-activated PDE is one of the most efficient enzymes [27]. As a consequence, cGMP hydrolysis by light-activated PDE is limited by the encounter rate $k_{\rm enc}$ between an activated PDE molecule diffusing on the disk surface and a cGMP molecule diffusing in the cytoplasms. Because cGMP diffusion in the cytoplasm is much faster than PDE diffusion in the membrane ($\sim 100\,\mu{\rm m}^2/{\rm s}^{-1}$ vs $0.8\,\mu{\rm m}^2/{\rm s}^{-1}$ [28], we can neglect PDE diffusion and assume that PDE is immobile. In that case, the mean rate hydrolysis rate is inversely proportional to the mean first passage time (MFPT) a cGMP molecule takes to find an activated PDE located on the surface of the compartment.

To estimate the encounter rate in a cylindrical compartment of radius R and height h, we compute the first passage time of a diffusing cGMP molecule to hit a circular spot of radius a located on the surface (Fig. 3a). The radius a equals the reaction radius between activated PDE and cGMP. To derive analytic expressions, we place the activated PDE molecule at the disk center. This assumption will not much affect the leading order term, because in a two or three-dimensional space, the leading order term of the mean first passage time to a small surface target does not depend on the position of target [29]. The first passage time $\tau(r,z)$ (cylindrical coordinates with rotational symmetry) of a cGMP molecule initially at position (r,z) satisfies the mixed boundary value problem [30]

$$\begin{aligned} D_g\Delta\tau(r,z) &= -1\,, \quad 0<z<h\,, 0\le r<R \\ \tau(r,z) &= 0\,, \quad z=0\,, r\le a \\ \frac{\partial}{\partial z}\tau(r,z) &= 0\,, \quad z=0\,, r>a \text{ and } z=h \\ \frac{\partial}{\partial r}\tau(r,z) &= 0 \quad r=R \end{aligned} \tag{43}$$

D_g is the cGMP diffusion coefficient. To obtain the MFPT $\bar{\tau}$, we average the solution $\tau(r,z)$ over a uniform initial distribution. The result is [25]

$$k_{\text{enc}} = \frac{1}{\bar{\tau}} = \frac{D_g}{R^2}\left(\pi\frac{h}{a}\frac{a_0(h/a)}{\sqrt{2}} + \frac{4\ln(R/a)-3}{8}\right)^{-1}, \tag{44}$$

where the function $a_0\left(\frac{h}{a}\right)/\sqrt{2} \in [0.07, 0.25]$ is shown in Fig. 2a of [14]. When $h \sim R$, $a_0(\frac{h}{a}) \sim \frac{1}{4}$. We note that the log-contribution in expression originates from the degenerated geometry (flat cylinder). An analytic closed representation of the function a_0 is unknown.

We recall that general expression of the MFPT $\bar{\tau}$ depends on the initial position in the dimensionless variables

$$x = \frac{r}{a}, \quad y = \frac{z}{a}, \quad \alpha = \frac{R}{a}, \quad \beta = \frac{h}{a}, \quad |\Omega| = \frac{|V|}{a^3} = \pi\beta\alpha^2,$$

is

$$\hat{\tau}(x,y) = \begin{cases} \displaystyle\sum_{n=0}^{\infty} b_n \frac{I_0(l_n x)}{I_0(l_n)} v_n(y) + w_i(x,y), & x \le 1 \\ \displaystyle = \sum_{n=0}^{\infty} a_n \frac{F_0(k_n x, k_n\alpha)}{F_0(k_n, k_n\alpha)} u_n(y) + \frac{\ln x}{2\pi\beta} - \frac{x^2-1}{4|\Omega|}, & 1 \le x \le \alpha, \end{cases} \tag{45}$$

where

$$u_0 = \frac{1}{\sqrt{2}}, \quad u_n(y) = \cos(k_n y) \quad (n \ge 1), \quad v_n(y) = \sin(l_n y) \quad (n \ge 0), \\ k_n = \frac{n\pi}{\beta}, \quad l_n = \frac{(n+\frac{1}{2})\pi}{\beta} \tag{46}$$

and the modified Bessel functions are $I_0(x)$ and $K_0(x)$ and the relations [31] ($I_0'(x) = I_1(x)$, $K_0'(x) = -K_1(x)$, we obtain

$$p_n(x) = \frac{F_0(k_n x, k_n\alpha)}{F_0(k_n, k_n\alpha)}, \tag{47}$$

with

$$F_0(x,y) = I_0(x)K_1(y) + K_0(x)I_1(y).$$

We recall that

$$\begin{aligned} w_i(x,y) &= \frac{1}{|\Omega|}\sum_{n=1}^{\infty} c_n J_0(z_n x)\frac{\cosh(z_n(\beta-y))}{\cosh(z_n\beta)} - \frac{x^2-1}{4|\Omega|} \\ &= \frac{1}{|\Omega|}\sum_{n=1}^{\infty} c_n J_0(z_n x)\left(\frac{\cosh(z_n(\beta-y))}{\cosh(z_n\beta)} - 1\right), \end{aligned} \tag{48}$$

where z_n are the positive zeros of the Bessel function $J_0(x)$, and the coefficients c_n are given by

$$c_n = \frac{2}{J_0'(z_n)^2}\int_0^1 J_0(z_n x)\frac{x^2-1}{4}x dx\,. \tag{49}$$

With the relation

$$\alpha_0 = 0\,, \quad \alpha_n = -k_n\frac{F_1(k_n,k_n\alpha)}{F_0(k_n,k_n\alpha)}\ (n\geq 1)\,, \quad \beta_n = l_n\frac{I_1(l_n)}{I_0(l_n)} \tag{50}$$

we obtain a closed matrix equations for the coefficient a_n and b_n

$$\begin{aligned} \sum_{m=0}^{\infty}(\beta_n+\alpha_m)\xi_{nm}a_m &= \sum_{m=0}^{\infty}\xi_{nm}\gamma_m \\ \sum_{m=0}^{\infty}(\beta_m+\alpha_n)\xi_{mn}b_m &= \gamma_n\,. \end{aligned} \tag{51}$$

where

$$\xi_{nm} = \begin{cases} \dfrac{2}{\beta}\dfrac{l_n}{l_n^2-k_m^2} = \dfrac{2}{\pi}\dfrac{(n+\frac{1}{2})}{(n+\frac{1}{2})^2-m^2}\,, & m\geq 1 \\ \dfrac{\sqrt{2}}{\beta l_n} = \dfrac{\sqrt{2}}{\pi}\dfrac{1}{n+\frac{1}{2}}\,, & m=0 \end{cases} \tag{52}$$

is an orthogonal matrix. It is possible to solve these matrix equations by truncating the system at a certain n leading to an approximated solution for a_n resp. b_n. This will lead to an approximation for the NET $\hat{\tau}(x,y)$ [14].

The encounter rate is given by relation (44) and clarifies how it depends on the underlying geometrical and diffusion properties. For example, for a toad rod with $R = 3\,\mu\text{m}$ we compute $k_{\text{enc}} \approx 2.9\,\text{s}^{-1}$, and for a mouse with $R = 0.7\,\mu\text{m}$ we find $k_{\text{enc}} \approx 61\,\text{s}^{-1}$ (with $a = 3\,\text{nm}$, $h = 15\,\text{nm}$, and $a_0(h/a) \approx 0.7$). As it turn out, the dependency of k_{enc} on the OS geometry is crucial to understand how many activated PDE is necessary to generate a signal that overcomes the noise. For mouse, the calculated rate $k_{\text{enc}} = 61\,\text{s}^{-1}$ for mouse is close to the value $43\,\text{s}^{-1}$ extracted from experimental data [26].

4 Modeling the Dynamics of cGMP and Calcium Ions

4.1 Coarse-Grained Model for cGMP Dynamics

This part of the model consists in first considering separately the dynamics occurring in each compartment: synthesis and hydrolysis of cGMP and second to couple cGMP to neighboring compartments through the reduced diffusion derived in Sect. 2.6.

The biochemistry can be described as follows: cGMP synthesis is catalyzed by guanylyl cyclase (GC) that are uniformly distributed on the surface of the disks. The synthesis rate depends on calcium through Ca^{2+}-sensitive guanylyl cyclase activating proteins (GCAPs) that inhibit GC at high Ca^{2+} concentration [32–35]. The calcium dependent cGMP synthesis rate in compartment n is described by the function

$$\alpha_s(n,t) = \alpha_{\max}\left(r_\alpha + (1-r_\alpha)\frac{K_\alpha^{n_\alpha}}{K_\alpha^{n_\alpha} + c(n,t)^{n_\alpha}}\right), \tag{53}$$

where $c(n,t)$ is the free Ca^{2+} concentration in compartment n, $\alpha_{\max}$ is the maximal synthesis rate for low free calcium, $r_\alpha = \frac{\alpha_{\min}}{\alpha_{\max}}$ is the ratio between minimal and maximal synthesis rate, K_α is the calcium concentration for which the synthesis rate is $(\alpha_{\max} + \alpha_{\min})/2$, and n_α is the Hill coefficient.

The rate of cGMP hydrolysis depends on the number of spontaneously activated PDE $P^*_{\rm sp}(n,t)$ and the number of light-activated PDE $P^*_{\rm li}(n,t)$,

$$\alpha_h(n,t) = k_{\rm sp}P^*_{\rm sp}(n,t) + k_{\rm li}P^*_{\rm li}(n,t)\,. \tag{54}$$

$k_{\rm sp}$ is the rate constant for a single spontaneously activated PDE, and $k_{\rm li}$ the diffusion limited rate constant for light-activated PDE, for which we use the encounter rate computed in Eq. (44).

The longitudinal cGMP diffusion between compartments occurs through the effective longitudinal diffusion constant $D_{g,l} < D_g$, where D_g is the fast cytosolic diffusion constant [20]. By applying Fick's law to model the longitudinal flux between neighboring compartments separated by the distance $h + w$ (compartment height plus disk width) we get the discrete flux

$$j_{d,g}(n,t) = \frac{D_{g,l}}{h(h+w)}(g(n+1,t) + g(n-1,t) - 2g(n,t))\,, \tag{55}$$

where $g(n,t)$ is the cGMP concentration in compartment n. Finally, the dynamics of cGMP across the ROS satisfies the equation

$$\frac{d}{dt}g(n,t) = j_{d,g}(n,t) + \alpha(n,t) - \left(k_{\rm sp}P^*_{\rm sp}(n,t) + k_{\rm li}P^*_{\rm li}(n,t)\right)g(n,t). \tag{56}$$

4.2 Reduced Model for Calcium Dynamics

To model Ca^{2+} dynamic, we take into account the effective longitudinal diffusion between compartments, the exchange between the OS and the extracellular medium through cGMP gated channels and $Ca^{2+}Na^{+}K^{+}$ exchangers, and the buffering activity. The model is presented now.

4.3 Modeling Calcium Buffers

In darkness, there a steady-state concentration of free calcium $c_d \approx 0.3\,\mu M$ [36] that corresponds on average to $\sim$3.3 ions in a compartment. This number is surprisingly small, given that many feedback process is regulated by Ca^{2+}. However, there are many Ca^{2+} binding proteins in the OS that contribute to buffer calcium and increase the amount of Ca^{2+} present in the OS, e.g., recoverin, GCAPs, and calmodulin. For example, the concentration of recoverin in a mammalian rod is $\sim 600\,\mu M$ [36], around 2000 times larger than the free calcium concentration; and the GC membrane concentration $\sim 50\,\mu m^{-2}$ [36] corresponds to $\sim$150 enzymes in a mouse compartment, around 40 times more than the number of free calcium ions.

In that model, we use the simplest buffering scenario: the buffering activity is much faster than the time scale where the free Ca^{2+} concentration fluctuates, and we use a linear relation between buffered and free Ca^{2+}, valid if the amount of buffered Ca^{2+} is small compared to the total buffer capacity. Hence, we consider that the number of bound calcium is

$$c_b(n,t) = B_{\mathrm{Ca}} c(n,t) \tag{57}$$

with the buffering capacity is B_{Ca}.

4.4 Dynamics of Calcium Exchange via Channels and Exchangers

Free internal Ca^{2+} ions are exchanged between the OS and the extracellular medium through cGMP gated channels and $Ca^{2+}Na^{+}K^{+}$ exchangers. The Ca^{2+} influx through the CNG channels depends on the probability $p_{\mathrm{ch}}(n,t)$ that a channel is open, which is a function of the local cGMP concentration:

$$p_{\mathrm{ch}}(n,t) = \frac{g(n,t)^{n_{\mathrm{ch}}}}{g(n,t)^{n_{\mathrm{ch}}} + K_{\mathrm{ch}}^{n_{\mathrm{ch}}}}\,. \tag{58}$$

The Ca^{2+}-efflux through exchangers depends on the free concentration $c(n,t)$ and the exchanger saturation level, leading to the relation

$$p_{\rm ex}(n,t) = \frac{c(n,t)}{c(n,t) + K_{\rm ex}}. \tag{59}$$

The net local Ca^{2+} membrane flux is

$$J_{\rm Ca}(n,t) = J_{\rm ch,Ca}(n,t) + J_{\rm ex,Ca}(n,t) = J_{\rm ch,Ca,max} p_{\rm ch}(n,t) + J_{\rm ex,Ca,max} p_{\rm ex}(n,t)\,. \tag{60}$$

The inward current through the CNG channels is carried by both ions Na^+ and Ca^{2+}

$$I_{\rm ch}(n,t) = I_{\rm ch,na}(n,t) + I_{\rm ch,Ca}(n,t) = \frac{I_{\rm ch,Ca}(n,t)}{f_{\rm Ca}} = -\frac{2\mathcal{F} J_{\rm ch,Ca}(n,t)}{f_{\rm Ca}}\,, \tag{61}$$

where $\mathcal{F}$ is the Faraday constant. There is only a fraction $f_{\rm Ca} \sim 0.1$–0.15 of the channel current carried by Ca^{2+} ions [36]. The extrusion of a single Ca^{2+} ion by the exchanger is accompanied by the influx of four Na^+ ions and the efflux of one K^+ [37]. Thus, the extrusion of one Ca^{2+} leads to the influx of a single positive charge, producing the net exchanger current

$$I_{\rm ex}(n,t) = \mathcal{F} J_{\rm ex,Ca}(n,t)\,. \tag{62}$$

Using Eqs. (61) and (62), we obtain for the local current

$$I(n,t) = I_{\rm ch}(n,t) + I_{\rm ex}(n,t) = \mathcal{F}\left(-\frac{2}{f_{\rm Ca}} J_{\rm ch,Ca}(n,t) + J_{\rm ex,Ca}(n,t)\right) \tag{63}$$

(see below for the analytical expressions). At steady-state in darkness, the calcium influx and efflux are balanced thus

$$J_{\rm ch,Ca}(n) + J_{\rm ex,Ca}(n) = 0. \tag{64}$$

From Eq. (63), we obtain the expression for the dark current associated with a single compartment

$$I_{{\rm comp},d} = -\mathcal{F}\left(\frac{2}{f_{\rm Ca}} + 1\right) J_{\rm ch,Ca,comp,d} = \mathcal{F}\left(\frac{2}{f_{\rm Ca}} + 1\right) J_{\rm ex,Ca,comp,d}\,. \tag{65}$$

We can use this result to express the calcium fluxes as a function of $I_{{\rm comp},d}$,

$$J_{\mathrm{ch,Ca}}(n,t) = J_{\mathrm{ch,Ca,comp},d}\frac{J_{\mathrm{ch,Ca}}(n,t)}{J_{\mathrm{ch,Ca,comp},d}} = -\frac{I_{\mathrm{comp},d}}{\mathcal{F}}\frac{f_{\mathrm{Ca}}}{f_{\mathrm{Ca}}+2}\frac{p_{\mathrm{ch}}(n,t)}{p_{\mathrm{ch},d}}$$
$$J_{\mathrm{ex,Ca}}(n,t) = J_{\mathrm{ex,Ca,comp},d}\frac{J_{\mathrm{ex,Ca}}(n,t)}{J_{\mathrm{ex,Ca,comp},d}} = \frac{I_{\mathrm{comp},d}}{\mathcal{F}}\frac{f_{\mathrm{Ca}}}{f_{\mathrm{Ca}}+2}\frac{p_{\mathrm{ex}}(n,t)}{p_{\mathrm{ex},d}}, \tag{66}$$

where g_d and c_d are the mean concentrations in darkness and

$$p_{\mathrm{ch},d} = \frac{g_d^{n_{\mathrm{ch}}}}{g_d^{n_{\mathrm{ch}}} + K_{\mathrm{ch}}^{n_{\mathrm{ch}}}} \quad \text{and} \quad p_{\mathrm{ex},d} = \frac{c_d}{c_d + K_{\mathrm{ex}}}. \tag{67}$$

Using Eq. (66) in Eq. (60) we obtain

$$J_{\mathrm{Ca}}(n,t) = V_{\mathrm{comp}}\phi\left(\frac{p_{\mathrm{ch}}(t)}{p_{\mathrm{ch},d}} - \frac{p_{\mathrm{ex}}(t)}{p_{\mathrm{ex},d}}\right), \tag{68}$$

where we use the notation to connect to each compartment: $I_{\mathrm{os},d} = N_{\mathrm{comp}} I_{\mathrm{comp},d}$, $V_{\mathrm{os}} = N_{\mathrm{comp}} V_{\mathrm{comp}}$ and

$$\phi = \frac{f_{\mathrm{Ca}}}{f_{\mathrm{Ca}}+2}\frac{|I_{\mathrm{comp},d}|}{V_{\mathrm{comp}}\mathcal{F}} = \frac{f_{\mathrm{Ca}}}{f_{\mathrm{Ca}}+2}\frac{|I_{\mathrm{os},d}|}{V_{\mathrm{os}}\mathcal{F}}. \tag{69}$$

For example, in a mouse rod with a dark current $I_{\mathrm{os},d} = 16\,\mathrm{pA}$ and a cytosolic volume $V_{\mathrm{os}} \approx 18, \mu\mathrm{m}^3 = 18 \times 10^{-15} l$ [36] we find $\phi \approx 500\frac{\mu\mathrm{M}}{\mathrm{s}} = 0.5\frac{\mu\mathrm{M}}{\mathrm{ms}}$.

4.5 Mass-Action Equation for the Free Ca^{2+} Concentration

The longitudinal calcium diffusion proceeds with an effective diffusion constant $D_{c,l}$. By considering buffering and the diffusion exchanges, we obtain for the free calcium concentration the equation

$$\frac{d}{dt}c(n,t) = j_{d,c}(n,t) + \frac{\phi}{B_{\mathrm{Ca}}+1}\left(\frac{p_{\mathrm{ch}}(t)}{p_{\mathrm{ch},d}} - \frac{p_{\mathrm{ex}}(t)}{p_{\mathrm{ex},d}}\right). \tag{70}$$

with the exchange rate

$$j_{d,c}(n,t) = \frac{1}{B_{\mathrm{Ca}}+1}\frac{D_{c,l}}{h(h+w)}(c(n+1,t) + c(n-1,t) - 2c(n,t)), \tag{71}$$

4.6 Coupled System of Equations for cGMP and Calcium Currents

We scale the various quantities using the mean dark concentrations g_d and c_d

$$\begin{aligned} \hat{g}(n,t) &= \frac{g(n,t)}{g_d}\,, \quad \widehat{\mathrm{Ca}}(n,t) = \frac{c(n,t)}{c_d}\,, \\ k_\alpha &= \frac{K_\alpha}{c_d}\,, \quad k_{\mathrm{ex}} = \frac{K_{\mathrm{ex}}}{c_d}\,, \quad k_{\mathrm{ch}} = \frac{K_{\mathrm{ch}}}{g_d}\,. \end{aligned} \tag{72}$$

The equations for the scaled cGMP and Ca^{2+} concentrations are

$$\begin{aligned} \frac{d\hat{g}(n,t)}{dt} &= j_{d,g}(n-1,n,n+1,t) + \beta_d \frac{r_\alpha + (1-r_\alpha)\dfrac{k_\alpha^{n_\alpha}}{k_\alpha^{n_\alpha} + \hat{c}(n,t)^{n_\alpha}}}{r_\alpha + (1-r_\alpha)\dfrac{k_\alpha^{n_\alpha}}{k_\alpha^{n_\alpha}+1}} \\ &\quad - \left(k_{\mathrm{sp}} P^*_{\mathrm{sp}}(n,t) + k_{\mathrm{li}} P^*_{\mathrm{li}}(n,t)\right)\hat{g}(n,t) \\ \frac{d\hat{c}(n,t)}{dt} &= j_{d,c}(n-1,n,n+1,t) + \gamma_d \left(\frac{p_{\mathrm{ch}}(n,t)}{p_{\mathrm{ch},d}} - \frac{p_{\mathrm{ex}}(n,t)}{p_{\mathrm{ex},d}}\right). \end{aligned} \tag{73}$$

with

$$\begin{aligned} \beta_d &= k_{\mathrm{sp}} \bar{P}^*_{\mathrm{sp,comp}} \\ \frac{p_{\mathrm{ch}}(n,t)}{p_{\mathrm{ch},d}} &= \frac{1 + k_{\mathrm{ch}}^{n_{\mathrm{ch}}}}{\hat{g}(n,t)^{n_{\mathrm{ch}}} + k_{\mathrm{ch}}^{n_{\mathrm{ch}}}} \hat{g}(n,t)^{n_{\mathrm{ch}}} \\ \frac{p_{\mathrm{ex}}(n,t)}{p_{\mathrm{ex},d}} &= \frac{1 + k_{\mathrm{ex}}}{\hat{c}(n,t) + k_{\mathrm{ex}}} \hat{c}(n,t)\,, \\ \gamma_d &= \frac{1}{B_{\mathrm{Ca}} + 1} \frac{\phi}{c_d} = \frac{1}{B_{\mathrm{Ca}} + 1} \frac{f_{\mathrm{Ca}}}{f_{\mathrm{Ca}} + 2} \frac{|I_{\mathrm{os},d}|}{c_d V_{\mathrm{os}} \mathcal{F}}\,. \\ j_{d,g}(n-1,n,n+1,t) &= \frac{D_{g,l}}{h(h+w)} \left(\hat{g}(n+1,t) + \hat{g}(n-1,t) - 2\hat{g}(n,t)\right) \\ j_{d,c}(n-1,n,n+1,t) &= \frac{1}{B_{\mathrm{Ca}} + 1} \frac{D_{c,l}}{h(h+w)} \left(\hat{c}(n+1,t) + \hat{c}(n-1,t) - 2\hat{c}(n,t)\right) \end{aligned} \tag{74}$$

By inserting Eq. (66) into Eq. (63), we obtain the normalized current

$$\hat{I}(n,t) = \frac{I_{\mathrm{comp},d} - I(n,t)}{I_{\mathrm{comp},d}} = 1 - \frac{2}{f_{\mathrm{Ca}} + 2} \frac{p_{\mathrm{ch}}(n,t)}{p_{\mathrm{ch},d}} + \frac{f_{\mathrm{Ca}}}{f_{\mathrm{Ca}} + 2} \frac{p_{\mathrm{ex}}(n,t)}{p_{\mathrm{ex},d}}\,. \tag{75}$$

Table 2 Parameters for the photocurrent simulation

Parameter	Definition
N_{comp}	Number of compartments
R	OS radius
h	Compartment height
w	Disk width
a	Reaction radius for cGMP hydrolysis by an activated PDE molecule
k_{enc}	Encounter rate between a cGMP and an activated PDE molecule
k_{li}	Rate constant for cGMP hydrolysis by a light-activated PDE
k_{sp}	Rate constant for cGMP hydrolysis by a spontaneous activated PDE
	Determined from the equation $\beta_d = k_{\text{sp}}\bar{P}^*_{\text{sp,comp}}$
β_d	cGMP hydrolysis rate in the dark
g_d	cGMP concentration in the dark
c_d	Free calcium concentration in the dark
$I_{\text{os},d}$	OS current in the dark
f_{Ca}	Fraction of current carried by calcium
B_{Ca}	Buffering capacity for calcium
K_α	Michaelis constant for cGMP synthesis
K_{ch}	Michaelis constant for channel opening
K_{ex}	Michaelis constant for calcium exchanger
n_α	Hill coefficient for cGMP synthesis
r_α	Ratio of minimal to maximal cGMP synthesis rate
n_{ch}	Hill coefficient for channel opening
D_g	Radial cGMP diffusion constant
D_{Ca}	Radial calcium diffusion constant
$D_{g,l}$	Effective longitudinal cGMP diffusion constant
$D_{\text{Ca},l}$	Effective longitudinal calcium diffusion constant
γ_d	Rate for calcium exchange

The overall normalized current from N_{comp} compartments is

$$\hat{I}_{\text{os}}(t) = \frac{I_{\text{os},d} - I_{\text{os}}(t)}{I_{\text{os},d}} = 1 - \frac{1}{I_{\text{os},d}} \sum_{n=1}^{N_{\text{comp}}} I(n,t) = 1 - \frac{1}{N_{\text{comp}}} \sum_{n=1}^{N_{\text{comp}}} \hat{I}(n,t). \quad (76)$$

The parameter of the simulations are summarized in Table 2:

5 Stochastic Simulations of the Dark Noise and the Single Photon Response (SPR)

We now describe the simulation method of the SPR with dark noise[38]. For each compartment, we use the SSA algorithm [17] to generate spontaneously activated PDE $P^*_{\text{sp}}(n,t)$ from Poisson activation and deactivation rates ν_{sp} and μ_{sp}. To model

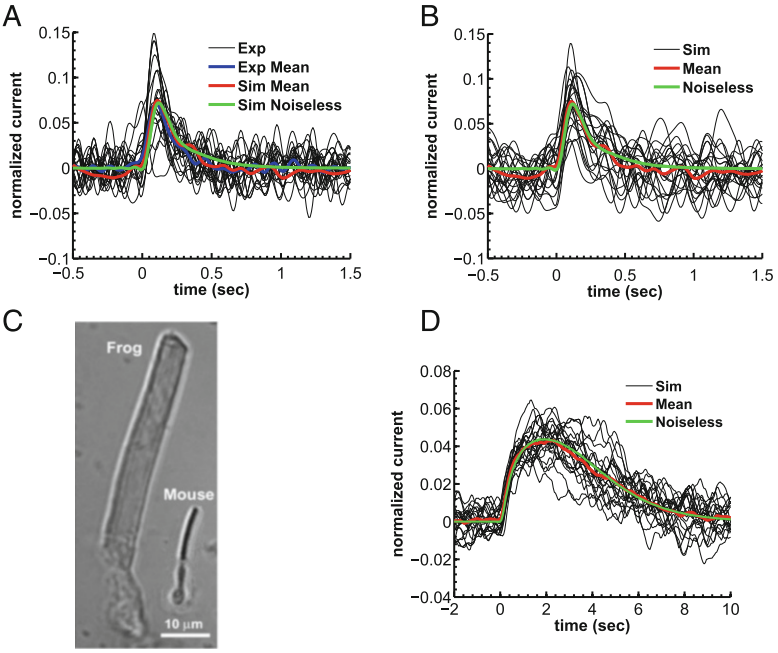


Fig. 4 Single photon responses of mouse and toad rods. (**a**) Electrophysiological recordings of single photon responses (*black*) from a mouse rod together with the mean response (*blue*). Mean (*red*) and the deterministic simulation (*green*) from **b** are superimposed for comparison. Currents have been normalized to the circulating current in darkness. (**b**) Single photon response simulations for a mouse rod (*black*) with mean (*red*) and a simulation of the mean response from a deterministic model without noise (*green*). (**c**) Rods from frog (toad) and mouse showing the large size difference [38]. (**d**) Single photon response simulations for a toad rod ($\beta_d = 1\,\mathrm{s}^{-1}$). Note the much larger duration compared to **b**

the single photon response, we simulated the number of light-activated PDE $P^*_{\mathrm{li}}(t)$ in the compartment where a photon is absorbed. For simplicity, we assume that a photon is absorbed at the center of the outer segment. However, a different location would not have a significant effect on the results. Finally, the functions $P^*_{\mathrm{sp}}(n,t)$ and $P^*_{\mathrm{li}}(t)$ are input to the system of equations for calcium and cGMP Eq. (73), from which we can compute the normalized currents $\hat{I}(n,t)$ and $\hat{I}_{\mathrm{os}}(t)$.

Combining all previous results into an integrated model, we can simulate a single photon response with intrinsic noise. We present 20 single photon responses (Fig. 4a) obtained from suction-electrode recordings in a WT mouse rod that we used to validate the model. Suction-electrode recording is used because cells could be held for longer times with this method, making it possible to obtain sufficient data from single cells over a period of several minutes. The data in Fig. 4a are representative of recordings from 8 rods, which all gave similar results. To generate the calculated single photon response curves shown in Fig. 4b, we used simulations of light-activated PDE from Fig. 2c.

Under the experimental recording conditions, the decay time of light-activated PDE is about 200 ms [39, 40], and the mean lifetime of excited rhodopsin is of the order of 40 ms [39, 41]. Furthermore, to reconcile the experimental and simulated response amplitude, we increased the transducin activation rates by a factor 1.75 compared to the toad simulations shown in Fig. 4b, which could be a result of the higher body temperature [42]. In addition, following this procedure, the average number of light-activated PDE increases from a value around 6 to around 8.2. The simulated responses in Fig. 4b show good agreement with the experimental recordings in Fig. 4a; however, the simulated dark noise ($\Sigma_{\text{sim}} = 2.3\%$) is higher compared to the recorded dark noise ($\Sigma_{\text{dark}} = 1.6\%$). A strong calcium feedback with no buffering ($B_{\text{Ca}} = 1$) and no saturation in cGMP synthesis at high calcium concentrations ($r_\alpha = 0$) reduces both the noise level and the peak amplitude by around 50% [38].

6 Statistical Analysis and Parameter Estimations

The current fluctuations in darkness (absence of photon) generate a noise called dark noise, the parameters of which can be extracted from the analytical expression of the power spectrum. We derive here such expression by considering that the dark noise is generated by the spontaneous activations and deactivations of PDE. From the expression of the dark noise, we estimate the spontaneous PDE activation process from electrophysiological recordings in the absence of photon response.

Using the model presented in the previous section, in darkness there is not light-activated PDE thus $P^*_{\text{li}}(n,t) = 0$. The average value of the scaled quantities CGMP and calcium $\hat{g}(n,t)$, $\hat{c}(n,t)$, and $\hat{I}(n,t)$ is one. Using a linear noise expansion of the Fourier transform of Eq. (73), we obtain

$$\delta\hat{c}(\omega) = \sum_{n=1}^{N_{\text{comp}}} \delta\hat{c}(n,\omega) = \frac{\gamma_d}{\gamma_d\xi_{\text{ex}} - i\omega}\xi_{\text{ch}}\delta\hat{g}(\omega)\,, \tag{77}$$

$$\delta\hat{g}(\omega) = \sum_{n=1}^{N_{\text{comp}}} \delta\hat{g}(n,\omega) = \frac{\beta_d}{\beta_d - i\omega - \beta_d\xi_\alpha \frac{\gamma_d\xi_{\text{ch}}}{\gamma_d\xi_{\text{ex}} - i\omega}} \sum_{n=1}^{N_{\text{comp}}} \delta\hat{P}^*_{\text{sp}}(n,\omega)\,. \tag{78}$$

with

$$\xi_\alpha = -n_\alpha \frac{1}{k_\alpha^{n_\alpha} + 1} \frac{(1-r_\alpha)\frac{k_\alpha^{n_\alpha}}{k_\alpha^{n_\alpha}+1}}{r_\alpha + (1-r_\alpha)\frac{k_\alpha^{n_\alpha}}{k_\alpha^{n_\alpha}+1}}\,, \quad \xi_{\text{ch}} = n_{\text{ch}} \frac{k_{\text{ch}}^{n_{\text{ch}}}}{1 + k_{\text{ch}}^{n_{\text{ch}}}}\,, \quad \xi_{\text{ex}} = \frac{k_{\text{ex}}}{1 + k_{\text{ex}}} \tag{79}$$

The overall current fluctuation (Eq. 76) is

$$\delta\hat{I}_{\rm os}(\omega) = \frac{1}{N_{\rm comp}}\left(\frac{2}{f_{\rm Ca}+2} + \frac{f_{\rm Ca}}{f_{\rm Ca}+2}\frac{\gamma_d\xi_{\rm ex}}{\gamma_d\xi_{\rm ex}-i\omega}\right)\xi_{\rm ch}\delta\hat{g}(\omega)$$
$$= \chi_I(\omega)\frac{1}{N_{\rm comp}}\sum_{n=1}^{N_{\rm comp}}\delta\hat{P}^*_{\rm sp}(n,\omega) \tag{80}$$

where the transfer function is defined by

$$\chi_I(\omega) = \left(1+\frac{f_{\rm Ca}}{f_{\rm Ca}+2}\frac{i\omega}{\gamma_d\xi_{\rm ex}-i\omega}\right)\frac{\xi_{\rm ch}\beta_d}{\beta_d\left(1-\frac{\gamma_d^2\xi_\alpha\xi_{\rm ch}\xi_{\rm ex}}{\gamma_d^2\xi_{\rm ex}^2+\omega^2}\right)-i\omega\left(1+\frac{\beta_d\gamma_d\xi_\alpha\xi_{\rm ch}}{\gamma_d^2\xi_{\rm ex}^2+\omega^2}\right)}$$
$$\approx -\frac{\xi_{\rm ch}\beta_d}{\beta_d\left(1-\frac{\gamma_d^2\xi_\alpha\xi_{\rm ch}\xi_{\rm ex}}{\gamma_d^2\xi_{\rm ex}^2+\omega^2}\right)-i\omega\left(1+\frac{\beta_d\gamma_d\xi_\alpha\xi_{\rm ch}}{\gamma_d^2\xi_{\rm ex}^2+\omega^2}\right)}. \tag{81}$$

Because PDE activations in different compartments are independent, the spectrum of the overall scaled current $\hat{I}_{\rm os}(t)$ is computed from the Lorentzian of a Poisson process (see Sect. 2.4):

$$S_{\hat{I}_{\rm os}}(\omega) = |\chi_I(\omega)|^2\frac{1}{N_{\rm comp}}\hat{S}_{P^*_{\rm sp}} = \frac{1}{N_{\rm comp}}\frac{|\chi_I(\omega)|^2}{P^*_{\rm sp,comp}}\frac{4\mu_{\rm sp}}{\mu_{\rm sp}^2+\omega^2} = \frac{|\chi_I(\omega)|^2}{P^*_{\rm sp,os}}\frac{4\mu_{\rm sp}}{\mu_{\rm sp}^2+\omega^2}\,. \tag{82}$$

The current variance is defined by

$$\Sigma_{\hat{I}_{\rm os}} = \frac{1}{2\pi}\int_0^\infty S_{\hat{I}_{\rm os}}(\omega)d\omega\,. \tag{83}$$

6.1 Power Spectrum and Variance for the Mutant GCAPs$^{-/-}$ Rod

In the mutant mice GCAPs$^{-/-}$, the Ca^{2+}-feedback on cGMP synthesis is abolished, which can be modeled by setting $n_\alpha = 0$ in Eq. (73). With $\xi_\alpha = 0$, the expression for $\chi_I(\omega)$ in Eq. (81) simplifies to

$$\chi_I(\omega) = -\frac{\xi_{\rm ch}\beta_d}{\beta_d - i\omega}\,. \tag{84}$$

In that case, the power spectrum and variance of the scaled current reduce to

$$S_{\hat{I}_{\rm os,gcap}}(\omega) = \frac{4\xi_{\rm ch}^2}{\bar{P}^*_{\rm sp,os}\mu_{\rm sp}}\frac{\beta_d^2\mu_{\rm sp}^2}{(\beta_d^2+\omega^2)(\mu_{\rm sp}^2+\omega^2)} \tag{85}$$

Table 3 Parameters used to simulate PDE activation

Parameter	Toad	Mouse
$\bar{P}^*_{\text{sp,comp}}$	1.25	0.9
$\bar{P}^*_{\text{li,max}}$	150	8.2
$\mu_{\text{sp}}(\text{s}^{-1})$	1.8	12.4
$\mu_p\ (\text{s}^{-1})$	0.625	5
$\tau_{\text{Rh}}\ (\text{s})$	3	0.04
N	6	6
$k_N\ (\text{s}^{-1})$	200	350
ω	0.1	0.1
$\mu_t\ (\text{s}^{-1})$	300	300

$$\Sigma_{\hat{I}_{\text{os,gcap}}} = \frac{\xi^2_{\text{ch}}}{\bar{P}^*_{\text{sp,os}}} \frac{1}{1+\frac{\mu_{\text{sp}}}{\beta_d}} = \frac{\xi^2_{\text{ch}}}{N_{\text{comp}}\bar{P}^*_{\text{sp,comp}}} \frac{1}{1+\frac{\mu_{\text{sp}}}{\beta_d}}\,. \tag{86}$$

With the parameter values for mouse given in Table 4, we obtain $\Sigma_{\hat{I}_{\text{os,gcap}}} \approx 0.055$, which agrees with the value 0.056 extracted from the GCAPs$^{-/-}$ simulations [38]. Various parameters are given in Table 3.

6.2 *Power Spectrum and Variance with Fast Calcium Dynamics*

The role of calcium feedback on cGMP synthesis is to reduce the current fluctuations. This feedback is efficient when the calcium dynamics is fast compared to the underlying PDE fluctuations, such that calcium changes can be used to monitor the PDE changes.

To estimate how much feedback reduces the current variance, we derive analytic expressions for the fast calcium dynamics that we compare to the ones for GCAPs$^{-/-}$ rods with no calcium feedback. For example, in a mouse rod, the rate constant γ_d governing the calcium dynamics in Eq. (73) has a value $\gamma_d \approx 1670\,\text{s}^{-1}$ (Eq. 74 with no buffering, $B_{\text{Ca}} = 1$). Adding buffers ($B_{\text{Ca}} > 1$) slows down the dynamics and reduces the feedback.

For $\gamma_d \gg \frac{\omega}{\xi_{\text{ex}}}$ and $\gamma_d \gg \beta_d\xi_\alpha\xi_{\text{ch}}$ Eq. (81) simplifies to

$$\chi_I(\omega) \approx -\frac{1}{\zeta}\frac{\xi_{\text{ch}}\tilde{\beta}_d}{\tilde{\beta}_d - i\omega} \quad \text{with} \quad \zeta = 1 - \frac{\xi_\alpha\xi_{\text{ch}}}{\xi_{\text{ex}}} \quad \text{and} \quad \tilde{\beta}_d = \beta_d\zeta\,. \tag{87}$$

The spectrum and variance of the dark noise with fast calcium dynamics is the product of two Lorenzians:

$$S_{\hat{I}_{\mathrm{os,fast\,Ca}}}(\omega) = \frac{1}{\zeta^2}\frac{4\xi_{\mathrm{ch}}^2}{\bar{P}^*_{\mathrm{sp,os}}\mu_{\mathrm{sp}}}\frac{\tilde{\beta}_d^2\mu_{\mathrm{sp}}^2}{(\tilde{\beta}_d^2+\omega^2)(\mu_{\mathrm{sp}}^2+\omega^2)} \tag{88}$$

$$\Sigma_{\hat{I}_{\mathrm{os,fast\,Ca}}} = \frac{1}{\zeta^2}\frac{\xi_{\mathrm{ch}}^2}{\bar{P}^*_{\mathrm{sp,os}}}\frac{1}{1+\frac{\mu_{\mathrm{sp}}}{\tilde{\beta}_d}}\,. \tag{89}$$

Compared to GCAPs$^{-/-}$ rods, calcium feedback reduces the amplitude of the dark noise by a factor

$$\rho = \sqrt{\frac{\Sigma_{\hat{I}_{\mathrm{os,gcap}}}}{\Sigma_{\hat{I}_{\mathrm{os\,fast\,Ca}}}}} = \zeta\sqrt{\frac{1+\frac{\mu_{\mathrm{sp}}}{\tilde{\beta}_d}}{1+\frac{\mu_{\mathrm{sp}}}{\beta_d}}}\,. \tag{90}$$

With mouse parameters from Table 4 we obtain $\rho \approx 2.5$ with $B_{\mathrm{Ca}} = 80$ and $r_\alpha = 0.066$, and $\rho \approx 4.4$ with strong feedback achieved for $B_{\mathrm{Ca}} = 0$ and $r_\alpha = 0$. From experimental recordings and dark noise simulations shown in Figs. 4 and 5, we find $\rho = 0.056/0.023 \approx 2.4$, in agreement with the theoretical value.

Table 4 Parameter values used to simulate the photocurrent

Parameter	Toad	Mouse
N_{comp}	2000	810
R (μm)	3	0.7
h (nm)	15	15
w (nm)	15	15
a (nm)	3	3
k_{enc}(s^{-1})	2.9	61
k_{li}(s^{-1})	2.9	61
β_d (s^{-1})	1	4.1
g_d (μM)	3	3
c_d (μM)	0.3	0.3
$I_{\mathrm{os},d}$ (pA)	40	17.9
f_{Ca}	0.12	0.12
γ_d (s^{-1})	92	23.4
B_{Ca}	1	80
K_α (μM)	0.15	0.1
K_{ch} (μM)	20	20
K_{ex} (μM)	1.6	1.6
n_α	2	2
r_α	0	0.066
n_{ch}	3	3
D_g (μm^2 s^{-1})	150	150
$D_{g,l}$ (μm^2 s^{-1})	20	40
$D_{\mathrm{Ca,l}}$ (μm^2 s^{-1})	20	20

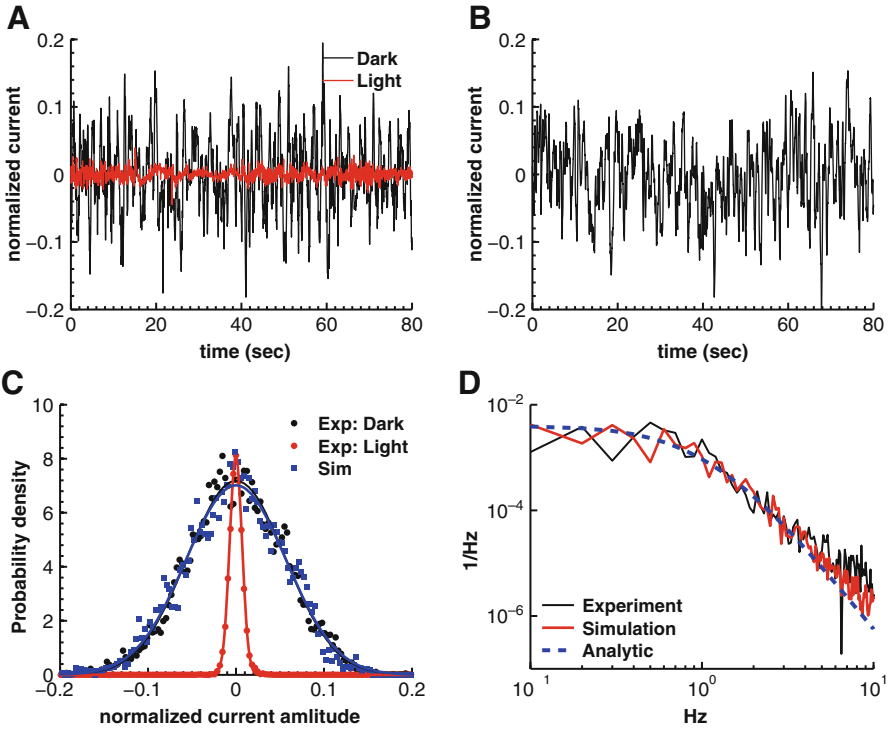


Fig. 5 Dark noise recordings and simulations for a GCAPs$^{-/-}$ mouse rod. (**a**) Electrophysiological current recordings in complete darkness (*black*) and in bright light (*red*). Bright light recordings are used to extract the instrumental noise. (**b**) Simulations of the current fluctuations in darkness (dark noise). (**c**) Probability distribution of the amplitudes from panels (**a**) and (**b**) together with Gaussian fits. (**d**) Comparison of the dark-light power spectrum from **a** with the power spectrum from **b** and analytic result

7 Parameter Extraction from Dark Noise Recordings

Experimental recordings of dark noise in wild type (WT) and GCAPs$^{-/-}$ knockout mice together can be used with expressions for the power spectrum and variance to evaluate unknown parameters in vivo.

7.1 Estimation of μ_{sp} and $P^{*}_{\mathrm{sp,comp}}$ from Dark Noise Recordings in GCAPs$^{-/-}$ Mice

For GCAPs$^{-/-}$ rods, the power spectrum divided by the variance [Eqs. (85) and (86)] reduces to a double Lorentzian that depends only on the parameters μ_{sp} and β_d:

$$\hat{S}_{\hat{I}_{\rm os}}(\omega) = 4\frac{\left(\beta_d + \mu_{\rm sp}\right)\beta_d\mu_{\rm sp}}{(\beta_d^2 + \omega^2)(\mu_{\rm sp}^2 + \omega^2)} \tag{91}$$

Because $\beta = 4.1\,\mathrm{s}^{-1}$ [26] is known for a mouse rod, we used Eq. (91) to extract the unknown spontaneous PDE deactivation rate $\mu_{\rm sp}$.

To extract $\mu_{\rm sp}$, we used the current recordings from GCAPs$^{-/-}$ mouse rods recorded in darkness and bright light conditions (Fig. 4a). The latter is needed to estimate the instrumental noise, since in bright light all channels are closed and the recorded noise is only instrumental noise [7]. Because instrumental and biological noise are independent, the dark noise power spectrum and variance can be computed by subtracting the instrumental values. Using Eq. (91) to fit the dark-light power spectrum scaled by the dark-light variance, it is possible to obtain an averaged value $\mu_{\rm sp} = 12.4\,\mathrm{s}^{-1}$. Subsequently, with the values of $\mu_{\rm sp}$ and β_d, we used Eq. (86) and fitted the unknown mean number of spontaneously activated PDE per compartment $\bar{P}^*_{\rm sp,comp}$ using the measured dark-light variance (with $N_{\rm comp} = 810$). We obtained an average value $\bar{P}^*_{\rm sp,comp} = 0.9$.

7.2 *Estimation of the Parameters r_α and $B_{\rm Ca}$ from Dark Noise Recordings in WT Mice*

The single photon response and the dark noise amplitude strongly depend on the calcium feedback. Equation (90) shows that the dark noise amplitude can be reduced by a factor 4.4 due to calcium feedback. By analyzing experimental data from WT and GCAPs$^{-/-}$ mice, we found a factor around 2.4 (Figs. 4 and 5). The strength of the calcium feedback depends on r_α and $B_{\rm Ca}$ (feedback on cGMP synthesis and buffering capacity). Unfortunately, both values are not precisely known. Most models assume $r_\alpha = 0$ [36, 43, 44], in [45] a value $r_\alpha = 0.072$ is used. In [36, 43] a buffering capacity $B_{\rm Ca} = 50$ is assumed, $B_{\rm Ca} = 20$ is used in [45] and $B_{\rm Ca} = 100$ in [44].

To estimate r_α and $B_{\rm Ca}$, we computed the dark-light power spectrum from dark noise recordings in WT rods (Fig. 5a). We fitted r_α and γ_d using Eq. (82) and then used Eq. (74) to compute $B_{\rm Ca}$ from γ_d. By fitting the spectrum we obtained $r_\alpha = 0.066$ and $\gamma_d = 23.4\,\mathrm{s}^{-1}$. With the experimental mean dark current of 17.9 pA, we then computed $B_{\rm Ca} = 80$, which is in agreement with experimental recordings [34, 46].

We used these values to simulate the dark noise in a WT rod (Fig. 4b). We quantified the agreement between data and simulations by comparing the probability distributions of the recorded and simulated current amplitudes (Fig. 4c), and by comparing the experimental dark-light spectrum with the spectrum extracted from the simulations and with the analytical expression in Eq. (82) (Fig. 5d). Although we find very good agreement for the power spectra (Fig. 5d), the standard deviation of the simulated current amplitude ($\Sigma_{\rm sim} = 2.3\%$) is about 15% smaller than the experimental value ($\Sigma_{\rm dark} = 2.7\%$). This difference may result from instrumental

noise that increases the recorded noise in darkness, which is not accounted for in the simulation. This effect is much larger for WT than GCAPs$^{-/-}$ rods because WT rods have less intrinsic dark noise.

8 Conclusion

Thirty years of modeling of single photon response connected to the statistical analysis of electrophysiological recordings led recently to the conclusion that the biochemistry and geometry of the rod may have evolved and adapted together to insure single photon detection across species [18, 38], but it is unclear whether these adaptations occurred independently or were coupled together by some other mechanism. This adaptation reveals that smaller rods are not a scaling copy larger one.

To conclude, the remarkable sensitivity of rods to single photons reveals a selection principle of evolution: an increase in the expression level of PDE compensate for the reduction in outer-segment geometry. Mouse rod can respond to a single photon by closing approximately the same percentage of outer segment channels as in a toad, but it can use many fewer G proteins and effector molecules and achieves higher temporal resolution. How the biochemistry of transduction and the geometry of the outer segment may have evolved together to ensure the detection of single photons is certainly a new question to address. The surprise of these researches is that the conclusion about the co-evolution of the cell geometry and the biochemistry came from the development of stochastic modeling of the underlying molecular processes. Similar modeling is expected in many other transduction processes such as heat sensing, olfaction, auditory transduction. The diversity of the cell geometries involved in transduction, that have evolved for billions of years, remains an open question for modern geometry, but the physiology needs to be taken into account.

References

1. K.-W. Yau, R. Hardie, Phototransduction motifs and variations. Cell **139**, 246–264 (2009)
2. Z. Song, M. Postma, S. Billings, D. Coca, R. Hardie, M. Juusola, Stochastic, adaptive sampling of information by microvilli in fly photoreceptors. Curr. Biol. **22**(15), 1371–1380 (2012)
3. G. Fain, in *Sensory Transduction* (Sinauer, Sunderland, 2003)
4. V. Arshavsky, M. Burns, Photoreceptor signaling: supporting vision across a wide range of light intensities. J. Biol. Chem. 287, 1620–1626 (2012)
5. D. Baylor, T. Lamb, K.-W. Yau, Responses of retinal rods to single-photons. J. Physiol. 288, 613–634 (1979)
6. G. Field, F. Rieke, Mechanisms regulating variability of the single-photon responses of mammalian rod photoreceptors. Neuron, 35, 733–747 (2002)
7. F. Rieke, D. Baylor, Molecular origin of continuous dark noise in rod photoreceptors. Biophys. J. **71**, 2553–2572 (1996)

8. S. Nikonov, T. Lamb, E. Pugh Jr, The role of steady phosphodiesterase activity in the kinetics and sensitivity of the light-adapted salamander rod photoresponse. J. Gen. Physiol. **116**, 795–824 (2000)
9. D.A. Baylor, G. Matthews, K.-W. Yau, Two components of electrical dark noise in toad retinal rod outer segments. J. Physiol. **309**, 591–621 (1980)
10. F. Rieke, D. Baylor, Single-photon detection by rod cells of the retina. Rev. Mod. Phys. **70**(3), 1027–1036 (1998)
11. G. Whitlock, T. Lamb, Variablity in the time course of single photon responses from toad rods: termination of rhodopsin's activity. Neuron **23**, 337–351 (1999)
12. D. Holcman, J. Korenbrot, The limit of photoreceptor sensitivity: molecular mechanism of dark noise in retinal cones. J. Gen. Physiol. **125**, 641–660 (2005)
13. T. Doan, A. Mendez, P. Detwiler, J. Chen, F. Rieke, Multiple phosphorylation sites confer reproducibility of the rod's single-photon responses. Science **13**, 530–533 (2006)
14. J. Reingruber, D. Holcman, The narrow escape problem in a flat cylindrical microdomain with application to diffusion in the synaptic cleft. SIAM Multiscale Model. Simul. **9**, 793–816 (2011)
15. J. Reingruber, D. Holcman, The dynamics of phosphodiesterase activation in rods and cones. Biophys. J. **94**(6), 1954–1970 (2008)
16. R. Hamer, S. Nicholas, D. Tranchina, P. Liebman, T. Lamb, Multiple steps of phosphorylation of activated rhodopsin can account for the reproducibility of vertebrate rod single-photon responses. J. Gen. Physiol. **122**, 419–444 (2003)
17. D.T. Gillespie, General method for numerically simulating stochastic time evolution of coupled chemical-reactions. J. Comp. Phys. **22**, 403–434 (1976)
18. J. Reingruber, G. Fain, D. Holcman, How rods respond to single photons: key adaptations of a G-protein cascade that enable vision at the physical limit of perception. Bioessays **37**(11), 1243–1252 (2015)
19. J. Reingruber, D. Holcman, Estimating the rate of cGMP hydrolysis by phosphodiesterase in photoreceptors. J. Chem. Phys. **129**, 145192 (2008)
20. D. Holcman, J. Korenbrot, Longitudinal diffusion in retinal rod and cone outer segment cytoplasm: the consequence of the cell structure. Biophys. J. **86**, 2566–2582 (2004)
21. D. Holcman, Z. Schuss, The narrow escape problem. SIAM Rev. **56**(2), 213–257 (2014)
22. D. Holcman, Z. Schuss, in *Stochastic Narrow Escape in Molecular and Cellular Biology: Analysis and Applications* (Springer, Berlin, 2015)
23. S. McLaughlin, J. Brown, Diffusion of calcium ions in retinal rods. A theoretical calculation. J. Gen. Physiol. **77**(4), 475–487 (1981)
24. D. Andreucci, P. Bisegna, G. Caruso, H. Hamm, E. DiBenedetto, Mathematical model of spatio-temporal dynamics of second messengers in visual transduction. Biophys. J. **85**, 1358–1376 (2003)
25. J. Reingruber, D. Holcman, Diffusion in narrow domains and application to phototransduction. Phys. Rev. E **79**(3), 030904 (2009)
26. O. Gross, E.N. Pugh, M. Burns, Spatiotemporal cgmp dynamics in living mouse rods. Biophys. J. **102**(8), 1775–1784 (2012)
27. I. Leskov, V. Klenchin, J. Handy, G. Whitlock, V. Govardovskii, M. Bownds, T. Lamb, E. Pugh Jr, V. Arshavsky, The gain of rod phototransduction: reconciliation of biochemical and electrophysiological measurements. Neuron **27**, 525–537 (2000)
28. Y. Koutalos, K. Nakatani, K.-W. Yau, Cyclic GMP diffusion coefficient in rod photoreceptors outer segment. Biophys. J. **68**, pp. 373–382 (1995)
29. A. Singer, Z. Schuss, D. Holcman, B. Eisenberg, Narrow escape I. J. Stat. Phys. **122**(3), 437–536 (2006)
30. Z. Schuss, *Theory and Applications of Stochastic Differential Equations*. Wiley Series in Probability and Statistics (Wiley, New York, 1980)
31. H. Carslaw, J. Jaeger, in *Conduction of Heat in Solids*, 2nd edn. (Oxford University Press, Oxford, 1986)

32. M.E. Burns, A. Mendez, J. Chen, D.A. Baylor, Dynamics of cyclic GMP synthesis in retinal rods. Neuron **36**, 81–91 (2002)
33. C. Makino, X. Wen, E. Olshevskaya, I. Peshenko, A. Savchenko, A. Dizhoor, Enzymatic relay mechanism stimulates cyclic GMP synthesis in rod photoresponse: biochemical and physiological study in guanylyl cyclase activating protein 1 knockout mice. PLoS One **7**(10), e47637 (2012)
34. A. Dizhoor, E. Olshevskaya, I. Peshenko, Mg2+/Ca2+ cation binding cycle of guanylyl cyclase activating proteins (GCAPs): role in regulation of photoreceptor guanylyl cyclase. Mol. Cell. Biochem. **334**, 117–124 (2010)
35. A. Mendez, M. Burns, I. Sokal, A. Dizhoor, W. Baehr, K. Palczewski, D. Baylor, J. Chen, Role of guanylate cyclase-activating proteins (GCAPs) in setting the flash sensitivity of rod photoreceptors. Proc. Natl. Acad. Sci. U. S. A. **98**, 9948–9953 (2001)
36. E. Pugh Jr T. Lamb, Phototransduction in vertebrate rods and cones: molecular mechanism of amplification, recovery and light adaptation, in *Handbook of Biological Physics*, vol. 3 (Elsevier, Amsterdam, 2000), pp. 183–255
37. L. Cervetto, L. Lagnado, R. Perry, D. Robinson, P. McNaughton, Extrusion of calcium from rod outer segments is driven by both sodium and potassium gradients. Nature **337**, 740–743 (1989)
38. J. Reingruber, J. Pahlberg, M. Woodruff, A. Sampath, G. Fain, D. Holcman, Detection of single photons by rod photoreceptors. Proc. Natl. Acad. Sci. U. S. A. **110**(48), 19378–19383 (2013)
39. C. Chen, M. Woodruff, F. Chen, D. Chen, G. Fain, Background light produces a recoverin-dependent modulation of activated-rhodopsin lifetime in mouse rods. J. Neurosci. **30**, 1213–1220 (2010)
40. M. Woodruff, K. Janisch, I. Peshenko, A. Dizhoor, S. Tsang, G. Fain, Modulation of phosphodiesterase6 turnoff during background illumination in mouse rod photoreceptors. J. Neurosci. **28**, 2064–2074 (2008)
41. M. Burns, E. Pugh Jr., Rgs9 concentration matters in rod phototransduction. Biophys. J. **16**, 1538–1547 (2009)
42. M. Heck, K. Hofmann, Maximal rate and nucleotide dependence of rhodopsin-catalyzed transducin activation: initial rate analysis based on a double displacement mechanism. J. Biol. Chem. **276**, 10000–10009 (2001)
43. O. Gross, E.N. Pugh, M. Burns, Calcium feedback to cGMP synthesis strongly attenuates single-photon responses driven by long rhodopsin lifetimes. Neuron **76**(2), 370–382 (2012)
44. R. Hamer, S. Nicholas, D. Tranchina, T. Lamb, J. Jarvinen, Toward a unified model of vertebrate rod phototransduction. Vis. Neurosci. **22** , 417–436 (2005)
45. G. Caruso, P. Bisegna, D. Andreucci, L. Lenoci, V. Gurevich, H. Hamm, E. DiBenedetto, Identification of key factors that reduce the variability of the single photon response. Proc. Natl. Acad. Sci. U. S. A. **108**, 7804–7807 (2011)
46. I. Peshenko, A. Dizhoor, Ca2+ and Mg2+ binding properties of GCAP-1. Evidence that Mg2+-bound form is the physiological activator of photoreceptor guanylyl cyclase. J. Biol. Chem. **281**(33), 23830–23841 (2006)

A Phenomenological Spatial Model for Macro-Ecological Patterns in Species-Rich Ecosystems

Fabio Peruzzo and Sandro Azaele

1 Introduction

In recent years important contributions to our understanding of community assembly and spatial ecology have come from the study of ecological patterns across scales [6, 16, 22, 28, 32]. Macroecology has been prolific at suggesting a wealth of interesting patterns and mechanisms [8].

For instance, considerable effort has been spent in understanding patterns such as the Relative Species Abundance (RSA)—which gives the probability of finding a species with n individuals living on a specific area. The RSA has a pivotal role in identifying the drivers of commonness and rarity in species-rich ecosystems, including tropical forests and coral reefs [5, 34–36], and has multi-faceted implications, including conservation strategies. This has stimulated a number of approaches attempting to explain the mechanisms underpinning the RSA curve, and there is an ongoing debate over the relative superiority of the proposed models without producing, however, a conclusive answer [21, 22]. So far, one of the main issues has been that many reasonable models are able to match empirical data fairly well, thereby hampering the possibility to support a particular theory. This suggests that we should prefer a model over another one, depending on its ability to produce multiple predictions—in addition to the original pattern—in agreement with empirical data and without any further parameter fitting. In many cases, authors have tried to explain empirical RSAs by means of stochastic, mean-field models which assume well-mixed populations [3, 7, 13], which usually are not.

In contrast, spatial aspects of biodiversity have been described by the so-called β-diversity, which overtakes the assumption of individuals placed uniformly at random

F. Peruzzo • S. Azaele (✉)
Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds LS2 9JT, UK
e-mail: S.Azaele@leeds.ac.uk

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7_15

in space by capturing key aspects of the spatial distribution of species, such as their characteristic spatial turnover. Indeed, classical approaches to population ecology have commonly overlooked the empirical finding that real populations are spatially clustered across a wide range of scales. However, spatial aggregation is important because it increases the turnover of species in space and therefore decreases the similarity of communities that are farther apart [4, 27]. One of the simplest ways to capture this similarity decay with spatial separation is to introduce the Pair Correlation Function (PCF) [5, 23], which can be defined—as we will do in the following—as the correlation in species' abundances of a pair of samples at a given distance.

Finally, another empirical pattern that has received a remarkable attention and has a long history of research is the Species–Area Relationship (SAR) [2, 5, 16, 32]—which describes how the average number of species increases with the size of the sampled area. This is considered one of the most important and, probably, universal ecological patterns, although the understanding of the underlying mechanistic causes of the SAR curve have progressed slowly and only recently.

The macro-ecological patterns that we have described so far are not independent from one another. Theoretical ecologists have been developing an understanding of the relationships among these patterns, and there is a growing appreciation that such macro-ecological measures of biodiversity are inter-related in a deep way. Since Harte and colleagues [18] first suggested that it should be possible to estimate the SAR for a region by examining scattered point survey data, several models have emerged. Some of them are purely geometrical [31] or based on the application of the maximum entropy to ecology [17]; other studies have also reported the effects of particular biological traits on the shape of SAR [11]. Here, for the sake of simplicity and to make analytical progress, we will assume neutral population dynamics [1, 6, 19, 29, 30].

The neutral theory of biodiversity is a theoretical framework for ecological communities with one trophic level, i.e., for species which compete for the same pool of limited resources. Examples are plant species in a forest, breeding birds in a large geographical region, hoverflies living in certain landscapes or coral colonies thriving in warm and shallow waters reachable by sunlight. In the neutral approximation all individuals have the same chances to die or survive and their competition does not depend on the species they belong to. Besides, the population dynamics is assumed to be fundamentally stochastic. Therefore, from the neutral standpoint, individuals' stochastic dynamics is more important than species identity, when it comes to explaining empirical community patterns. However crude and unrealistic these assumptions may look like, they are at the core of models that are in good agreement with empirical measurements at stationarity. Despite such agreements do not necessarily imply that the population dynamics is neutral at the individual level, neutral theory is useful to unveil universal community patterns and it is, probably, more valuable when it fails than when it matches the data. Falsifying one or more of its assumptions, in fact, may inform key aspects of community dynamics.

In the following we will focus on a phenomenological neutral model, whose dynamics is spatially explicit and stochastic. Because it cannot be solved explicitly in full generality, we will introduce a method for calculating analytically approximate formulæ for the three patterns which we have alluded to above. We will then compare the analytical expressions with the numerical integration and, finally, we will show that the model is able to describe the empirical RSA, SAR and PCF of two tropical forests which harbour hundreds of plant species. With this model one can translate information from one pattern to another and extrapolate patterns outside the region of parametrization.

1.1 The RSA in the Mean-Field Approximation

Before introducing the spatial stochastic model, in order to make it clear how the neutral assumption enters the definition of a model, let us first focus on a simple form of RSA that can be deduced at the mean-field level. If we assume that the dynamics is Markovian and described by a birth and death (one-step) process, then in general the birth and death rates of species α can be written down as $b_\alpha(n_1, n_2, \ldots, n_S)$ and $d_\alpha(n_1, n_2, \ldots, n_S)$, respectively, where n_i is the population size of species i and S is the total number of species in a given region. If interactions are neutral, then those rates should be symmetric functions of species' population sizes and should not depend on the species label α (strictly speaking, this defines a symmetric model [6]—not a neutral one—but in the following we will not make such a distinction). Also, if we further assume that species are independent, then the birth and death rates factorize and we can focus on the dynamics of just one species, because any species is not affected by the presence of the others. In this way, the neutral and the independence assumptions allow us to think of the population sizes of species as independent realizations of a stochastic process. In our case, the birth and death rates are b_n and d_n, respectively, with n the number of individuals of a species in a given region. Therefore, the time evolution of the probability distribution of n is described by the following master equation:

$$\frac{\partial p_n(t)}{\partial t} = p_{n+1}(t)\, d_{n+1} + p_{n-1}(t)\, b_{n-1} - p_n(t)\, (b_n + d_n), \tag{1}$$

where $p_n(t)$ is the probability that a species has n individuals at time t. Of course, this equation needs to be equipped with boundary conditions that prevent n from becoming negative. Common choices are either reflecting or absorbing boundaries, depending on the nature of the problem. When $n = 0$ is reflecting, the equilibrium solution can be easily calculated [33] and is, for $n > 0$,

$$P_n = P_0 \prod_{i=0}^{n-1} \frac{b_i}{d_{i+1}}, \tag{2}$$

where P_0 is a normalization constant. If individuals belonging to abundant and rare species have the same chances to die, or survive and give birth to an offspring, then the per capita birth and death rates cannot depend on n and therefore, for $n \geq 0$, we have to set

$$b_n = gn + \delta_{n,0}\nu \qquad d_n = rn,$$

where g and r are positive constants, and ν is the speciation rate. In this framework there is no explicit biological mechanism for speciation: ν is a parameter that ensures that the system is always populated by one individual whenever species go extinct (reflecting boundaries). Let's denote by Φ_n the number of species with n of individuals. If S is the empirical number of species in our ecosystem, from Eq. (2) we get

$$\langle \Phi_n \rangle = SP_0 \prod_{i=0}^{n-1} \frac{b_i}{d_{i+1}} = SP_0 \frac{b_0 b_1 \dots b_{n-1}}{d_1 d_2 \dots d_n} = \theta \frac{x^n}{n}, \tag{3}$$

where $x = g/r < 1$, $n > 0$ and $\theta = SP_0\nu/g$ is known as the biodiversity parameter. Equation (3) is known amongst ecologists as 'Fisher log-series', and was first discovered experimentally in 1943 [14]. This distribution has no internal mode and therefore it predicts that singleton species (i.e., those with one individual only) are always the most frequent. This is not always the case, as many communities have species' abundances that are more frequent than singletons. These RSAs can be more adequately explained with an alternative choice of rates, i.e.,

$$b_n = gn + b \qquad d_n = rn, \tag{4}$$

where the parameter $b > 0$ incorporates immigration. Ultimately, in this setting rare species have a mild reproductive advantage over the more common ones. The equilibrium solution is the following negative binomial distribution:

$$\langle \Phi_n \rangle = S(1-x)^{\frac{b}{g}} \left(\frac{b}{g}\right)_n \frac{x^n}{n!}, \tag{5}$$

where $(a)_n = a(a+1)\dots(a+n-1)$ with $(a)_0 = 1, n = 0, 1, \dots$ and $x = g/r$ with $0 < x < 1$. This distribution can produce an internal mode in species' abundances and predicts that communities should harbour only a few species that are common and many species that are rare. This RSA is more flexible than the Fisher log-series and is in good agreement with empirical data [3, 36].

1.2 A Mean-Field Langevin Equation for the RSA

Larger areas of species-rich communities often sustain larger populations and support more species because, typically, they encompass greater habitat diversity and richer pool of resources. This simple observation shows that community patterns at relatively large spatial scales might be described by models which treat population size as a continuous random variable. Also, it suggests to include the principal effects driving the macro-ecological patterns in a simplified, phenomenological fashion. Within the neutral approach and assuming that the effects we outlined in the previous section are the most important driving factors, we get the following Fokker–Planck (FP) equation for the diffusive approximation of the master equation (Eq. (1)) with rates defined in Eq. (4)

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n}\Big[(b-\mu n)\,P(n,t)\Big] + \sigma^2 \frac{\partial^2}{\partial n^2}\Big[(n+\epsilon)\,P(n,t)\Big], \tag{6}$$

where $\mu = r - g > 0$, $\sigma^2 = (r+g)/2$ and $\epsilon = b/(r+g) > 0$. The equilibrium solution of this equation provides the continuous RSA, i.e.,

$$P(n) = P_0 (n+\epsilon)^{\frac{b+\mu\epsilon}{\sigma^2}-1} e^{-\frac{\mu n}{\sigma^2}}, \tag{7}$$

where P_0 is a normalization constant. A given large region is usually affected by a small immigration rate, which hence suggests that ϵ is typically a small parameter. If we treat it as such, then Eq. (7) can be approximated (at zeroth order) by a (normalized) gamma distribution of the form:

$$P(n) = \left(\frac{\mu}{\sigma^2}\right)^{\frac{b}{\sigma^2}} \frac{n^{\frac{b}{\sigma^2}-1} e^{-\frac{\mu n}{\sigma^2}}}{\Gamma(b/\sigma^2)}, \tag{8}$$

where $\Gamma(x)$ is the gamma function. The (non-uniform) correction to this equation is of order $\epsilon \ln \epsilon$ for $b/\sigma^2 \geq 1$ and ϵ^{b/σ^2} for $0 < b/\sigma^2 < 1$. In real species-rich ecological communities one typically finds $r \simeq g$ (usually, $1 - g/r < 0.02$, hence μ is positive and small [35]), which therefore allows the existence of a few species with a large number of individuals (population sizes larger than $\sigma^2/\mu = r/(r-g)$ when $0 < b/\sigma^2 < 1$ and $\epsilon \ll 1$). Rare species, instead, have population sizes typically smaller than $b/\mu = b/(r-g)$ (for $0 < b/\sigma^2 < 1$ and $\epsilon \ll 1$). As expected, Eq. (8) is the equilibrium solution of the simpler FP equation

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n}\Big[(b-\mu n)\,P(n,t)\Big] + \sigma^2 \frac{\partial^2}{\partial n^2}\Big[nP(n,t)\Big], \tag{9}$$

which corresponds to the Langevin equation (in the Itō prescription)

$$\dot{n} = b - \mu n + \sigma\sqrt{n}\,\xi(t), \tag{10}$$

where $\xi(t)$ is a zero mean white noise with $\langle \xi(t)\xi(t') \rangle = 2\delta(t-t')$. Equation (10) has a nice interpretation: in the limit of a small immigration rate, the dynamics of the RSA results from the trade-off between net immigration and net death rates (i.e., $b - \mu n$), and the fluctuations about these deterministic terms are simply driven by the central limit theorem (i.e., fluctuations $\propto \sqrt{n}$). The agreement of Eq. (8) with the data [3], therefore, suggests that demographic stochasticity may play a major role in sculpting macro-ecological patterns, including the RSA. In the following section these considerations will form the backbone of the spatial version of the model, thus extending the importance of the effects of immigration, birth, death and demographic stochasticity to spatial patterns as well.

2 A Phenomenological Spatial Stochastic Model: Linking Macro-Ecological Patterns

The assumption of well-mixed populations, of course, cannot account for the spatial turnover of species and the increase of species richness with sampled area. These two patterns are captured by the PCF and SAR, respectively, as explained in the introduction. A region with a high rate of spatial turnover of species, in which the PCF decays steeply, has also a steep increase in the SAR, because a given area contains relatively more species compared to other regions where the PCF decays more gradually. Also, empirical data highlight that the PCF is, typically, a monotonically decreasing function of distance. This underlines the important role of spatial clumping of individuals, because if individuals were found somewhere, it would be more likely to find other ones close-by.

These observations lead naturally to a simple spatial extension of the continuous model of the RSA. Since we are interested in spatial patterns at relatively large scales, we consider a phenomenological generalization in which space is coarse grained. We assume space is partitioned by a mesh into a collection of voxels—or, more precisely, a regular graph (or lattice) in which each vertex has $2d$ nearest neighbours, being d space dimension. Within each voxel (or, equivalently, vertex or site, which hereinafter will be used as synonyms), individuals are considered well-mixed, diluted and treated as point-like particles which undergo the demographic dynamics defined by Eq. (10), which incorporates birth, death and immigration (in the language of chemical reaction kinetics, these are first-order reactions known as autocatalytic production, degradation and production from source, respectively).

As the customary approach in the reaction-diffusion master equation (RDME), we will assume that, within a hypercubic voxel of width a (a is the lattice spacing as well), individuals are uniformly placed at random in space (i.e., voxels have no internal spatial structure). Therefore, a should be much smaller than all the other macroscopic length scales of interest, including the characteristic spatial correlation length of the system. In the following numerical integration and empirical analysis,

this will always be the case. The set of coupled stochastic differential equations defining the model are

$$\dot{n}_i(t) = D\nabla_i^2 n_i(t) + b - \mu n_i(t) + \sigma\sqrt{n_i(t)}\,\xi_i(t), \tag{11}$$

where $n_i(t)$ is the density of individuals in the i-th site at time t, $\xi_i(t)$ is a zero mean white noise (depending on site i) with correlation $\langle \xi_i(t)\xi_j(t')\rangle = 2\delta(t-t')\delta_{i,j}$. D is the "diffusion" coefficient and

$$\nabla_i^2 n_i(t) = \frac{1}{a^2}\sum_{j\in\partial(i)}[n_j(t) - n_i(t)], \tag{12}$$

where $\partial(i)$ indicates the set of nearest neighbours of i. There is nothing special about our choice of local movement, more general connectivities could have been chosen to study the effects of different topologies on macroscopic patterns [24]. More importantly—and unlike the RDME approach—here individuals move locally on the mesh in a deterministic fashion, as governed by the discrete Laplacian. This is tantamount to neglect contributions to stochasticity due to the random hopping of individuals, which is expected to be a good approximation for large diffusion constants [9]. Therefore, linear reactions taking place inside voxels—independent of diffusion—are supposed to be the main source of stochasticity in the system. In this framework, individuals do not undergo a continuous time random walk on the mesh, as can be seen from Eq. (11) when the internal demographic dynamics is switched off. This is one of the main reasons why this spatial stochastic model, at least in the current formulation, cannot be considered an appropriate coarse-grained approximation of an underlying microscopic, spatially continuous model. However, these approximations are not expected to have large effects on the first two moments, which we will study in the following sections and are at the core of our analysis. This is only a phenomenological framework which provides an analytical way to calculate macro-ecological patterns, starting from simple yet important demographic and spatial factors. Yet, microscopic models which are continuous in space, such as independent branching Brownian processes (or superprocesses [12]), might probably have a discrete approximation close to the current formulation. This will be investigated in a future work.

If we indicate with $\{n\}$ a given configuration of population sizes on the lattice, i.e., $\{n\} = \{n_1, n_2, \ldots\}$, the probability density function of $\{n\}$, $P(\{n\})$, satisfies the following FP equation (*sensu* Itō):

$$\partial_t P(\{n\},t) = -\sum_z \frac{\partial}{\partial n_z}\Big[\Big(D\nabla_z^2 n_z(t) + b - \mu n_z\Big)P(\{n\},t)\Big] + \sigma^2\sum_z \frac{\partial^2}{\partial n_z^2}\Big[n_z P(\{n\},t)\Big], \tag{13}$$

where the sums are over all sites of the lattice. It is easy to see that the average density per site is $\langle n_i\rangle = b/\mu$. It is interesting to notice that this model has a non-trivial stationary distribution only for $b > 0$ and when the per capita death rate

is strictly larger than the per capita birth rate (i.e., $\mu > 0$), because of the lack of a carrying capacity. In this sense, it is a minimal model for calculating large scale patterns: if one sets to zero one or more parameters, then the predicted macro patterns—if they exist—are trivial.

2.1 Calculating the Pair Correlation Function

The PCF describes the correlation in species' population abundances between different spatial locations. As we mentioned before, it plays a crucial role in linking some of the most important macro-ecological patterns.

Let's consider two sites i and j in a (d-dim) lattice and calculate $\langle n_i n_j \rangle$. Multiplying Eq. (13) by $n_i n_j$ and integrating all n's from zero to infinity (or using the usual Itō formula with Eq. (11)), one finds the equation for the time evolution of $\langle n_i n_j \rangle$, i.e.,

$$\frac{\partial}{\partial t} \langle n_i n_j \rangle = D(\nabla_i^2 \langle n_i n_j \rangle + \nabla_j^2 \langle n_i n_j \rangle) + 2b \langle n \rangle - 2\mu \langle n_i n_j \rangle + 2\sigma^2 \langle n \rangle \delta_{ij}, \tag{14}$$

where $\langle n \rangle = b/\mu$ and δ_{ij} is the Kronecker delta. Because we are interested in stationary patterns, we drop the time derivative and simplify the equation by looking at the correlation $G_{i,j} = \langle n_i n_j \rangle - \langle n_i \rangle \langle n_j \rangle = \langle n_i n_j \rangle - \langle n \rangle^2$. $G_{i,j}$ actually satisfies

$$D(\nabla_i^2 G_{i,j} + \nabla_j^2 G_{i,j}) - 2\mu G_{i,j} + 2\sigma^2 \langle n \rangle \delta_{ij} = 0 \quad . \tag{15}$$

In order to solve this equation, let us introduce a system of Cartesian coordinates and indicate with $\mathbf{x}$ the d-dim position vector of a site. Basically, in the previous equation we make the substitution $i \to \mathbf{x}$ and $j \to \mathbf{y}$, with the agreement that changes in any direction in the coordinates have to be made in multiples of a, the lattice spacing. In this way, we can use Fourier series to find an expression for $G_{\mathbf{x},\mathbf{y}}$ in an infinite lattice. After some algebraic manipulations, we finally get

$$G_{\mathbf{x},\mathbf{y}} = \left(\frac{a}{2\pi}\right)^d \frac{\sigma^2 b}{\mu^2} \int_{\mathcal{C}} \mathrm{d}\mathbf{p} \, \frac{e^{i\mathbf{p}\cdot(\mathbf{x}-\mathbf{y})}}{1 + \frac{2D}{\mu a^2} \sum_{i=1}^{d} (1 - \cos(p_i a))}, \tag{16}$$

where p_i is the i-th Cartesian component of $\mathbf{p}$ and $\mathcal{C}$ is the hypercubic (d-dim) primitive unit cell with size $2\pi/a$. As expected, $G_{\mathbf{x},\mathbf{y}}$ is translational invariant and in $d = 1$ reduces to a simple exponential:

$$G_{x,y} = Ck^{|x-y|/a}, \tag{17}$$

where $x, y = 0, a, 2a, \ldots$; $k < 1$ and C are positive constants which can be either calculated from Eq. (16) or by direct substitution into Eq. (15). For k one gets

$$k = 1 + \frac{\mu a^2}{2D} - \sqrt{\frac{\mu^2 a^4}{4D^2} + \frac{\mu a^2}{D}}, \tag{18}$$

from which one deduces the correlation length $\xi = -a/\ln(k)$. Notice that $\xi \to \sqrt{D/\mu}$ when $a \to 0$.

Instead of trying to calculate explicitly the integral in Eq. (16), we can obtain a good deal of simplification and insight by taking its continuum spatial limit (i.e., $a \to 0$ and the parameters are appropriately re-defined). Such a limit leads to

$$\begin{aligned} \mathcal{G}(\mathbf{x}, \mathbf{y}) &= \frac{1}{(2\pi)^d} \frac{\sigma^2 b}{\mu^2} \int_{\mathbb{R}^d} \mathrm{d}\mathbf{p}\, \frac{e^{i\mathbf{p}\cdot(\mathbf{x}-\mathbf{y})}}{1 + \frac{D}{\mu}\mathbf{p}^2} \\ &= \frac{\hat{\rho}^2 \langle n \rangle^2}{\left(2\pi\hat{\lambda}^2\right)^{d/2}} \left(\frac{|\mathbf{x}-\mathbf{y}|}{\hat{\lambda}}\right)^{(2-d)/2} K_{(2-d)/2}\left(\frac{|\mathbf{x}-\mathbf{y}|}{\hat{\lambda}}\right), \end{aligned} \tag{19}$$

where $K_\nu(x)$ is the modified Bessel function of the second kind of order ν or Macdonald's function [20], $\mathbf{x}$ and $\mathbf{y}$ are now continuous vector coordinates and

$$\hat{\lambda} = \sqrt{\frac{D}{\mu}}, \quad \hat{\rho} = \sqrt{\frac{\sigma^2}{b}}$$

are constants with length dimension when $d = 2$. As expected, $\mathcal{G}(\mathbf{x}, \mathbf{y})$ is also the solution of the continuum spatial limit of Eq. (15) in Cartesian coordinates (and dimension d), i.e.,

$$D\nabla^2_{\mathbf{z}} \mathcal{G}(\mathbf{z}) - \mu \mathcal{G}(\mathbf{z}) + \sigma^2 \langle n \rangle \delta(\mathbf{z}) = 0, \tag{20}$$

where $\mathbf{z} = \mathbf{x} - \mathbf{y}$, $\delta(\mathbf{z})$ is a Dirac delta and we took advantage of the translational symmetry of the system.

Of course, $\mathcal{G}$ obtained in Eq. (19) may be a good approximation of the discrete correlation only for $|\mathbf{x} - \mathbf{y}| \gg a$. As a first approximation, however, one may introduce a lower cut-off to $\mathcal{G}$ by stipulating that $\mathcal{G}(\mathbf{z}) = G_{\mathbf{x},\mathbf{x}}$ for all $|\mathbf{z}| \leq a$. Because $K_\nu(x)$ decays exponentially fast for large x [20], Eq. (19) also suggests that $\hat{\lambda}$ is the spatial correlation length of the system. Therefore, this continuous framework works under the condition that $\hat{\lambda} \gg a$, which is always satisfied in the following analysis.

In the next sections we look into the stationary Pair Correlation Function (PCF) defined as

$$g_{\mathbf{x},\mathbf{y}} = \frac{\langle n_{\mathbf{x}} n_{\mathbf{y}} \rangle}{\langle n \rangle^2}, \tag{21}$$

because—in a first approximation—it allows one to study the empirical properties of $g_{\mathbf{x},\mathbf{y}}$ independently of a, the spatial resolution introduced to calculate the PCF from the data. As an analytic expression, we will use its continuous version, i.e., $g(\mathbf{x},\mathbf{y}) = 1 + \mathcal{G}(\mathbf{x},\mathbf{y})/\langle n\rangle^2$, where $\mathcal{G}(\mathbf{x},\mathbf{y})$ is given in Eq. (19) with $d = 2$. Hence, the PCF reduces to

$$g(r) = 1 + \frac{1}{2\pi}\left(\frac{\hat{\rho}}{\hat{\lambda}}\right)^2 K_0\left(\frac{r}{\hat{\lambda}}\right), \tag{22}$$

where $r = |\mathbf{x} - \mathbf{y}|$. We will always assume that r is much larger than a.

3 A Method for Calculating Macro-ecological Patterns

The model defined in Eq. (11) is linear and therefore all the stationary n-point correlation functions can be calculated explicitly. However, having all correlation functions is not sufficient, in general, to build up a closed-form solution of the model, from which one derives all interesting patterns.

The spatial Relative Species Abundance (sRSA) is defined as the probability that a species has n individuals within a certain area A, if there are S_0 species in total in the larger area A_0 where A is contained. Therefore, the sRSA is given by the conditional probability $p(n|A, \{S_0, A_0\})$, and all correlation functions contribute to such distribution in a non-trivial way. So, instead of trying to calculate the sRSA from the correlation functions or the generating functional, we introduce an approximation which allows to make some analytical progress. Afterwards, we will check with the numerical integration of the model that such approximations are good, at least in the region of the parameter space which is relevant to the empirical patterns.

Because the calculations turn out to be easier in the continuum space, in what follows we will essentially work with Eqs. (19–20), bearing in mind that the results in such limit have to be used *cum grano salis*. For simplicity then, let us focus on a circular region, C, of radius R and define the random variable

$$N(R) = \int_C n(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{23}$$

which gives the number of individuals of a species living on C at stationarity. Of course, $\langle N(R)\rangle = \langle n\rangle\, \pi R^2$, but we can also calculate the variance, $\mathrm{Var}(N(R))$. From Eq. (19) we get

$$\int_C \int_C \mathcal{G}(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = \left\langle N(R)^2\right\rangle - \langle N(R)\rangle^2 = \mathrm{Var}(N(R)) \tag{24}$$

and the final expression in $d = 2$ is

$$\mathrm{Var}(N(R)) = \langle n\rangle\hat{\rho}^2\langle N(R)\rangle\Big(1 - \frac{2\hat{\lambda}}{R}\frac{K_1(R/\hat{\lambda})I_1(R/\hat{\lambda})}{K_0(R/\hat{\lambda})I_1(R/\hat{\lambda}) + K_1(R/\hat{\lambda})I_0(R/\hat{\lambda})}\Big), \quad (25)$$

where $I_\nu(x), K_\nu(x)$ are modified Bessel functions of the first and second kind of order ν, respectively [20]. Of course, this formula is reliable only when $R \gg a$, but it is interesting to notice that for $R \gg \hat{\lambda}$ the variance to mean ratio tends to a constant, i.e.,

$$\frac{\mathrm{Var}(N(R))}{\langle N(R)\rangle} \simeq \langle n\rangle\hat{\rho}^2 = \frac{\sigma^2}{\mu}, \quad (26)$$

which is exactly the ratio one obtains from the mean-field model, i.e., Eq. (8). Therefore, at stationarity the system reaches non-Poissonian fluctuations and on large spatial scales it is homogenized by diffusion. This is an example of a result that can be proved under quite general conditions [15].

As we have alluded to above, a lot of species-rich ecological communities have per capita birth and death rates that are very close ($1 - g/r < 0.02$, hence μ is positive and small [35]). So, we can roughly estimate the variance to mean ratio as

$$\frac{\mathrm{Var}(N(R))}{\langle N(R)\rangle} \simeq \frac{r}{r-g} \gg 1, \quad (27)$$

for $R \gg \hat{\lambda}$. Moreover, when $r \simeq g$ both the correlation length, $\hat{\lambda}$, and the correlation time, μ^{-1}, of the system are very large. This depicts such empirical communities as they were posed close to a critical point, where large fluctuations have a long-time behaviour and are correlated across many spatial scales.

Along the lines we have outlined before, one could in principle write down the expressions for the higher moments of $N(R)$. However, a deeper insight and more analytical progress can be achieved by introducing the following crucial approximation: we assume that, at stationarity, the random variable $N(R)$ is distributed according to the probability density function defined in Eq. (8)—the equilibrium solution of the mean-field model—with appropriate scale-dependent functions, $\alpha(R)$ and $\beta(R)$, which we are going to introduce. This is tantamount to assume that the functional form of the sRSA is the same across all spatial scales and hence the dependence on the spatial scale of the sRSA comes only through such functions. We have borrowed this hypothesis from the phenomenological renormalization group [26].

In order for the gamma distribution in Eq. (8) to match the first two moments of $N(R)$ that we have calculated, we then introduce a shape function, $\alpha(R)$, and a scale function, $\beta(R)$, both depending on R. The final approximate sRSA, $q(N|R)$, has therefore the form

$$q(N|R) = \frac{1}{\beta(R)}\frac{(N/\beta(R))^{\alpha(R)-1}}{\Gamma(\alpha(R))}e^{-N/\beta(R)}, \quad (28)$$

where $\Gamma(x)$ is a gamma function. From the properties of the gamma distribution, it is not difficult to show that, if we choose

$$\alpha(R) = \left(\frac{\langle N(R)\rangle}{\sigma(R)}\right)^2 \quad \text{and} \quad \beta(R) = \frac{\sigma(R)^2}{\langle N(R)\rangle}, \tag{29}$$

then we match exactly the first two moments, $\langle N(R)\rangle$ and $\mathrm{Var}(N(R))$. We will show that the approximate expression for the sRSA is in good agreement with the numerical integration of the model. With the formula for $q(N|R)$ one can directly link the sRSA to the PCF. In fact, when fitting the PCF and obtaining $\langle n\rangle$ from the data, we can predict the distribution of species' population sizes across all spatial scales by using Eq. (28).

Also, since a species can be observed only when it has at least one individual, the probability that a species is present within an area of radius R is $\int_1^\infty q(N|R)\mathrm{d}N$, from which one can calculate the SAR, an important pattern in many applications.

4 Numerical Scheme for the Integration of the Model

Naïve numerical schemes for integrating Eq. (11) are affected by severe drawbacks. For instance, if we apply a first-order explicit Euler method to the simpler Eq. (10) (*sensu* Itō), we get

$$n(t+\Delta t) = n(t) + \Delta t[b - \mu n(t)] + \sigma\sqrt{\Delta t\, n(t)}N(0,1), \tag{30}$$

where N(0,1) is a zero mean normal random variable with variance 1. It is well known that, starting from $n(0) > 0$, this method inevitably leads to produce negative values for $n(t+\Delta t)$, especially when $n(t)$ is small. Such unphysical densities are even more harmful when integrating stochastic partial differential equations, strongly biasing spatial correlations.

Building on previous methods [10, 25], we introduce a numerical integration scheme which generates (in the weak sense) the field $n_i(t)$ at stationarity in 2-dim—the d-dim case is straightforward—and ensures, by construction, that the density is always non-negative.

We first write down the discrete Laplacian on a 2-dim lattice of mesh size a, where every site has four nearest neighbours. Secondly, we re-write Eq. (11) as

$$\dot{n}_{\mathbf{x}}(t) = Y_{\mathbf{x}}(t) - \Omega n_{\mathbf{x}}(t) + \sigma\sqrt{n_{\mathbf{x}}(t)}\,\xi_{\mathbf{x}}(t), \tag{31}$$

where

$$Y_{\mathbf{x}}(t) = \frac{D}{a^2}\sum_{i=1}^{4} n_{\mathbf{x}+a\mathbf{e}_i}(t) + b \quad \text{and} \quad \Omega = \frac{4D}{a^2} + \mu, \tag{32}$$

and $\mathbf{e}_1 = (1,0)$, $\mathbf{e}_2 = (-1,0)$, $\mathbf{e}_3 = (0,1)$ and $\mathbf{e}_4 = (0,-1)$.

The stationary solutions of the FP equations associated with each local Langevin equation for $n_{\mathbf{x}}(t)$, i.e., Eq. (31), are gamma distributions given by Eq. (8) in which, locally, $b \to Y_{\mathbf{x}}$ and $\mu \to \Omega$. $Y_{\mathbf{x}}$ is then a new immigration parameter which accounts for the global as well as the local influx of individuals from the 4 nearest neighbours into the site with $\mathbf{x}$ coordinates; Ω is a new death rate which includes the possibility that individuals leave the site at $\mathbf{x}$ because of diffusion, in addition to the demographic death rate. If we initialize the lattice with $n_{\mathbf{x}}^{(0)} \geq 0$, we can then update each and every site by sampling from the local gamma distribution, conditioning on the nearest neighbours. Hence, at the $m+1$ sampling step the local density is given by

$$n_{\mathbf{x}}^{(m+1)} = \text{Gamma}\left[\frac{Y_{\mathbf{x}}^{(m)}}{\sigma^2}, \frac{\sigma^2}{\Omega}\right], \tag{33}$$

where $m \in \mathbb{N}$,

$$Y_{\mathbf{x}}^{(m)} = \frac{D}{a^2}\sum_{i=1}^{4} n_{\mathbf{x}+a\mathbf{e}_i}^{(m)} + b \tag{34}$$

and Gamma$[\alpha, \beta]$ is a gamma variate with shape parameter α and scale parameter β. One keeps updating the system until all the stationary summary statistics of interest do not change significantly in different generations (or they match a stationary summary statistics calculated analytically from the model).

Because $n_{\mathbf{x}}^{(0)} \geq 0$ and also $Y_{\mathbf{x}}$ and Ω are strictly positive at all steps (D, b, μ and σ are all strictly positive), by construction $n_{\mathbf{x}}$ is always non-negative and finite at all steps.

4.1 Comparisons with Analytical Solutions

We implemented the numerical scheme on a 200×200 lattice with periodic boundary conditions. Each site was initialized by drawing from a gamma distribution with shape parameter $\alpha = b/\sigma^2$ and scale parameter $\beta = \sigma^2/\mu$. The comparisons between the analytical formulæ obtained in the continuum approximation and the numerical integrations were carried out by considering 1000 independent realizations at stationarity on the square lattice. The results for the numerical and analytical PCF are shown in Fig. 1 for the correlation length $\hat{\lambda} = 10$.

For a given realization at stationarity, we decided that a species is observable—that is, it has at least one individual—within a given area C of radius R, if $N(R) = \sum_{\mathbf{x}\in C} n_{\mathbf{x}} \geq 1$. This, of course, resembles what happens in empirical observations

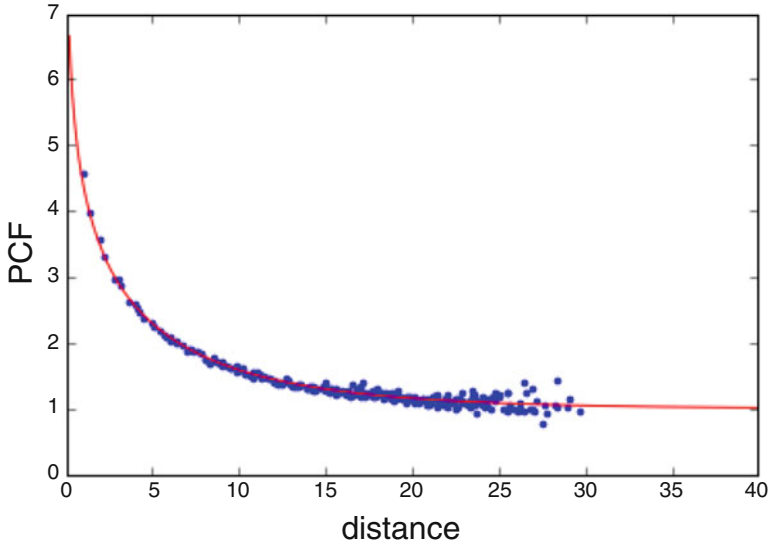


Fig. 1 Comparison between the analytical (see Eq. (22)) and numerical PCF calculated from the stationary densities generated by implementing the numerical scheme outlined in Sect. 4. Here the parameters are $D = 1$, $b = 0.005$, $\mu = 0.01$, $\sigma = 2.1$ and the distance is in lattice spacing units

and here we modify the previous definitions of sRSA and SAR by stipulating that a species can be observed only if it occurs with at least one individual. So, when an area of radius R_0 harbours $S(R_0)$ species in total, at smaller radii we define the sRSA as

$$\mathtt{sRSA}(R) = \frac{q(N|R)}{\int_1^\infty q(M|R_0)dM}, \tag{35}$$

where $q(N|R)$ is the distribution that we have obtained in Eq. (28). From this expression we can derive the SAR, which accounts for the number of species that are found within a certain area as a function of its radius. This is defined as

$$\mathtt{SAR}(R) = S(R_0)\frac{\int_1^\infty q(N|R)dN}{\int_1^\infty q(M|R_0)dM} \quad . \tag{36}$$

We have benchmarked the results for the sRSA and SAR obtained from the numerical scheme against the analytical formulæ in Figs. 2 and 3.

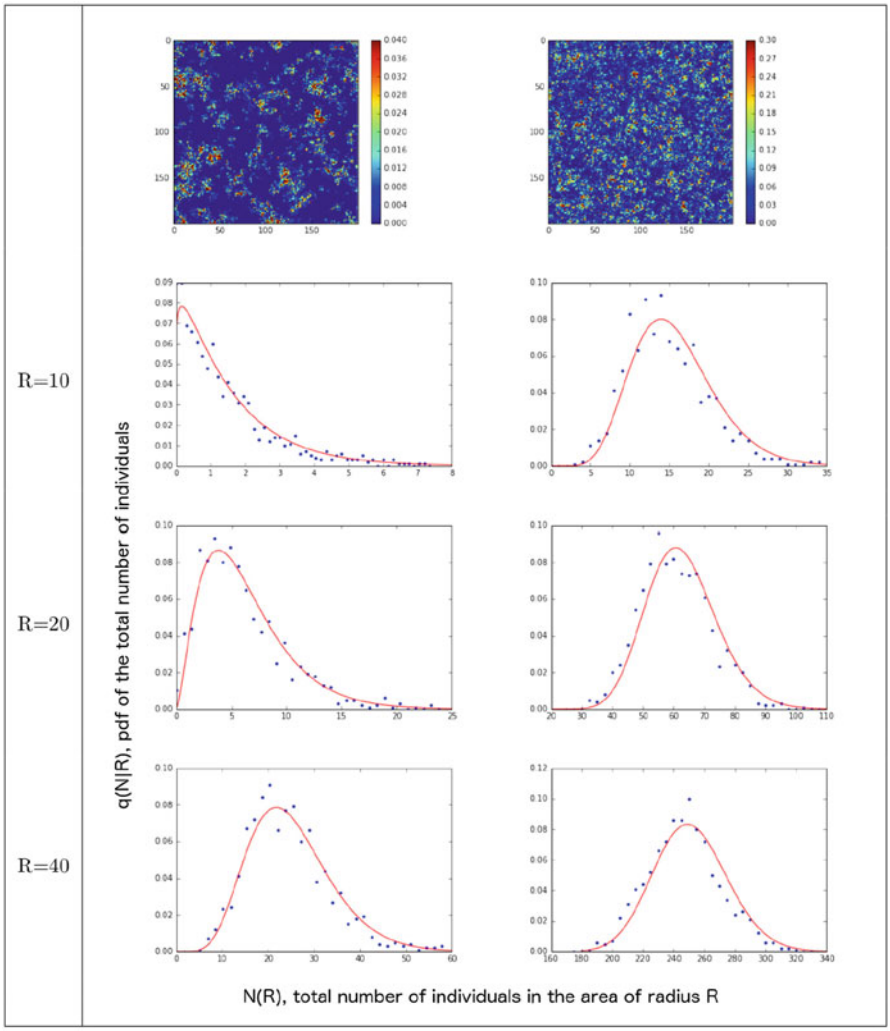


Fig. 2 Comparison between the analytical (see Eqs. (35), (28) and (29)) and numerical sRSA calculated from the stationary densities generated by implementing the numerical scheme outlined in Sect. 4. Here the parameters are $D = 100$, $b = 0.005$, $\mu = 1$, $\sigma = 2.1$ for the *left column* and $D = 1$, $b = 0.005$, $\mu = 0.1$, $\sigma = 0.5$ for the *right column*. The *two upper panels* depict two snapshots of the stationary densities on the corresponding lattices. The radius is in lattice spacing units

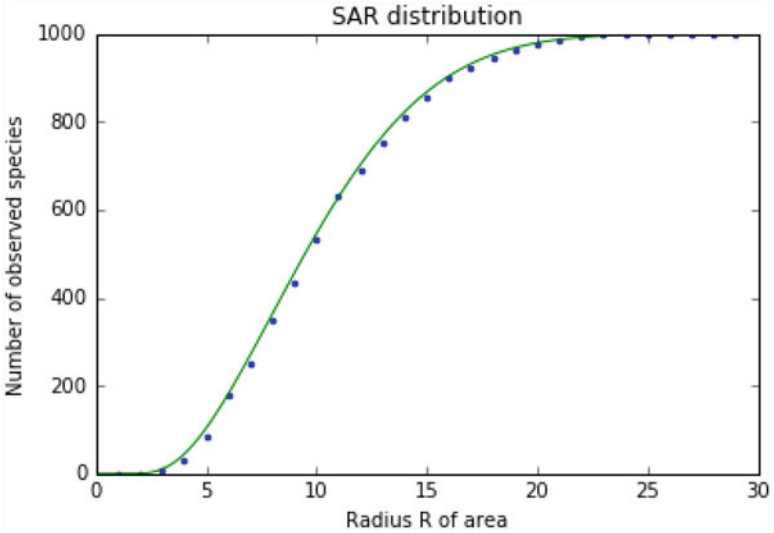


Fig. 3 Comparison between the analytical (see Eqs. (36), (28) and (29)) and numerical SAR calculated from the stationary densities generated by implementing the numerical scheme outlined in Sect. 4. Here the parameters are $D = 100$, $b = 0.005$, $\mu = 1$, $\sigma = 2.1$ and the radius is in lattice spacing units

5 Macro-ecological Patterns of Pasoh and Barro Colorado Island Forests

We considered two datasets from well-known forest stands: one set is from the Barro Colorado Island (BCI) in Panama and the other one from the Pasoh Forest Reserve in Malaysia. Both cover an area of 50 ha and were comprehensively surveyed, containing high but greatly different numbers of vascular plant species. Species identity, geographical location and diameter at breast height (DBH) were recorded for each tree living within the plot. We used such datasets of plant species to test model predictions against empirical patterns.

We first coarse grained the two systems by superimposing a grid mesh of 10m size and counted the number of individuals of each species within every sub-area. Then we looked at each pair of sites located at $\mathbf{x}$, $\mathbf{y}$ and calculated the empirical PCF with the following formula:

$$g_{\mathbf{x},\mathbf{y}} = \frac{\frac{1}{S}\sum_{\mu=1}^{S} n_{\mathbf{x}}^{(\mu)} n_{\mathbf{y}}^{(\mu)}}{(\frac{1}{S}\sum_{\mu=1}^{S} n_{\mathbf{x}}^{(\mu)})(\frac{1}{S}\sum_{\mu=1}^{S} n_{\mathbf{x}}^{(\mu)})} \tag{37}$$

where $n_{\mathbf{x}}^{(\mu)}$ is the number of individuals of species μ within the site located at $\mathbf{x}$ and S is the total number of species in the whole region. Then we calculated the parameters $\hat{\lambda}$ and $\hat{\rho}$ by best-fitting the data to the analytical formula in Eq. (22). Finally, from the empirical data we estimated $\langle n \rangle = N_0/(S_0 A_0)$ in both forests, where N_0 is the total number of individuals, S_0 is the total number of species in the whole area, A_0, of the forest plot. We found the ratio $\hat{\lambda}/\hat{\rho} \sim 0.33$ for Pasoh, and $\hat{\lambda}/\hat{\rho} \sim 0.35$ for BCI. These parameters are sufficient to predict the behaviour of the analytical SAR and sRSA with no further best-fit, and such predictions can therefore be compared to the empirical distributions for the two datasets. The agreement with empirical data is good as shown in Fig. 4.

6 Conclusions

We have introduced a phenomenological stochastic model, defined on a d-dim lattice, from which one can derive analytical approximations of important macro-ecological patterns, such as the PCF, the SAR and the sRSA. We devised an efficient numerical integration scheme, which confirms the goodness of the analytical derivations. Also, all the empirical patterns obtained from two canopy forests, the BCI and Pasoh plots, show a good agreement with the formulæ derived from the model, using three free parameters only. The framework is able to explain and link empirical macro-ecological patterns in a theoretically consistent way. Intriguingly, it suggests that many species-rich ecosystems may possibly be close to a critical point, in which slow and large fluctuations are correlated on large spatial scales. The theoretical setting calls for more refined spatial formulations and better articulated ecological mechanisms, which can provide more realism to the predictions as well as bridge the gap between individual behaviour and emergent macroscale patterns.

Acknowledgements The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme 'Stochastic Dynamical Systems in Biology: Numerical Methods and Applications' where work on this paper was undertaken. This work was supported by EPSRC grant no EP/K032208/1. We are also grateful to the FRIM Pasoh Research Committee (M.N.M. Yusoff, R. Kassim) and the Center for Tropical Research Science (R. Condit, S. Hubbell, R. Foster) for providing the empirical data of the Pasoh and BCI forests, respectively. SA is in debt with Prof. A. Maritan for insightful discussions.

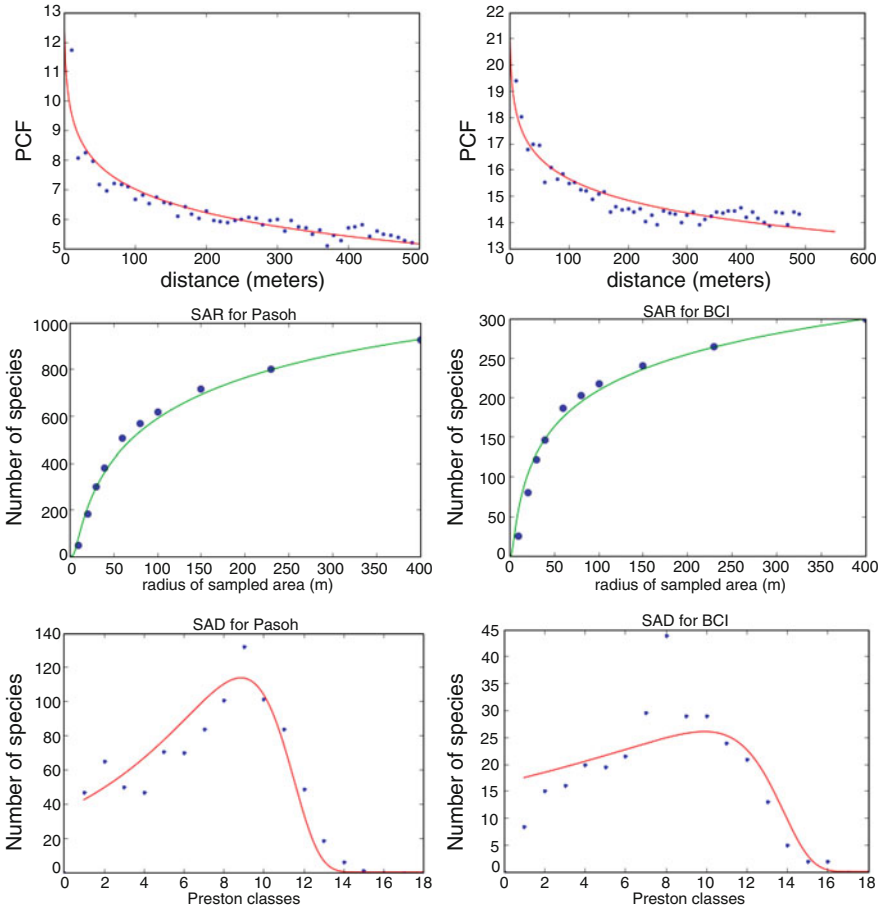


Fig. 4 The PCF, SAR and Species Abundance Distribution (SAD) for Pasoh (*left column*) and BCI (*right column*) tropical forests for trees that are larger than 10 cm in stem diameter at breast height. The first panel in each column shows the PCF from which we best-fitted the parameters $\hat{\lambda}$ and $\hat{\rho}$ (empirical data showed with *blue dots*). We found the ratio $\hat{\lambda}/\hat{\rho} \sim 0.33$ for Pasoh, and $\hat{\lambda}/\hat{\rho} \sim 0.35$ for BCI. The second panel depicts the SAR: *blue dots* are empirical data, *green line* is the predicted distribution by using the best-fitted parameters from the previous PCF, $\langle n \rangle$ and formulæ in Eqs. (36), (28) and (29). The third panel shows the SAD (this is defined as the sRSA times the total number of species in the region) for the whole area. The *blue dots* are empirical data, whereas the *red solid line* was obtained by using the best-fitted parameters from the previous PCF, $\langle n \rangle$ and formulæ in Eqs. (35), (28) and (29). Preston classes are customarily used in ecological studies and are similar to a $\log_2$-binning, although not exactly equivalent. Preston's binning method is described in Volkov et al. (2003)

References

1. D. Alonso, A.J. McKane, Sampling Hubbell's neutral theory of biodiversity. Ecol. Lett. **7**(10), 901–910 (2004)
2. O. Arrhenius, Species and area. J. Ecol. **9**(1), 95–99 (1921)
3. S. Azaele, S. Pigolotti, J.R. Banavar, A. Maritan, Dynamical evolution of ecosystems. Nature **444**(7121), 926–928 (2006)
4. S. Azaele, R. Muneepeerakul, A. Maritan, A. Rinaldo, I. Rodriguez-Iturbe, Predicting spatial similarity of freshwater fish biodiversity. Proc. Natl. Acad. Sci. **106**(17), 7058–7062 (2009)
5. S. Azaele, A. Maritan, S.J. Cornell, S. Suweis, J.R. Banavar, D. Gabriel, W.E. Kunin, Towards a unified descriptive theory for spatial ecology: predicting biodiversity patterns across spatial scales. Methods Ecol. Evol. **6**, 324–332 (2015)
6. S. Azaele, S. Suweis, J. Grilli, I. Volkov, J.R. Banavar, A. Maritan, Statistical mechanics of ecological systems: neutral theory and beyond. Rev. Mod. Phys. **88**(3), 035003 (2016)
7. A.J. Black, A.J. McKane, Stochastic formulation of ecological models and their applications. Trends Ecol. Evol. **27**(6), 337–345 (2012)
8. J.H. Brown, *Macroecology* (The University of Chicago Press, Chicago, 1995)
9. D.S. Dean, Langevin equation for the density of a system of interacting Langevin processes. J. Phys. A: Math. Gen. **29**(24), L613 (1996)
10. I. Dornic, H. Chaté, M.A. Munoz, Integration of Langevin equations with multiplicative noise and the viability of field theories for absorbing phase transitions. Phys. Rev. Lett. **94**(10), 100601 (2005)
11. S. Drakare, J.J. Lennon, H. Hillebrand, The imprint of the geographical, evolutionary and ecological context on species–area relationships. Ecol. Lett. **9**(2), 215–227 (2006)
12. A. Etheridge, *An Introduction to Superprocesses* (American Mathematical Society, Providence, 2000)
13. R.S. Etienne, D. Alonso, A.J. McKane, The zero-sum assumption in neutral biodiversity theory. J. Theor. Biol. **248**(3), 522–536 (2007)
14. A.S. Fisher, R.A. Corbet, C.B. Williams, The relation between the number of species of individuals in a random sample of an animal population. J. Anim. Ecol. **12**(12), 42–58 (1943)
15. C.W. Gardiner, M.L. Steyn-Ross, Adiabatic elimination in stochastic systems. II. Application to reaction diffusion and hydrodynamic-like systems. Phys. Rev. A **29**(5), 2823 (1984)
16. J. Grilli, S. Azaele, J.R. Banavar, A. Maritan, Spatial aggregation and the species–area relationship across scales. J. Theor. Biol. **313**, 87–97 (2012)
17. J. Harte, *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics* (Oxford University Press, Oxford, 2011)
18. J. Harte, S. McCarthy, K. Taylor, A. Kinzig, M.L. Fischer, Estimating species-area relationships from plot to landscape scale using species spatial-turnover data. Oikos **86**, 45–54 (1999)
19. S. Hubbell, *The Unified Theory of Biodiversity and Biogeography* (Princeton University Press, Princeton, 2001)
20. N.N. Lebedev, *Special Functions and Their Applications* (Courier Dover Publications, New York, 1972)
21. F. May, T. Wiegand, S. Lehmann, A. Huth, Do abundance distributions and species aggregation correctly predict macroecological biodiversity patterns in tropical forests? Glob. Ecol. Biogeogr. **25**, 575–585 (2016)
22. B.J. McGill, R.S. Etienne, J.S. Gray, D. Alonso, M.J. Anderson, H.K. Benecha, M. Dornelas, B.J. Enquist, J.L. Green, F. He, Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol. Lett. **10**(10), 995–1015 (2007)
23. H. Morlon, G. Chuyong, R. Condit, S. Hubbell, D. Kenfack, D. Thomas, R. Valencia, J.L. Green, A general framework for the distance–decay of similarity in ecological communities. Ecol. Lett. **11**(9), 904–917 (2008)

24. R. Muneepeerakul, S. Azaele, S.A. Levin, A. Rinaldo, I. Rodriguez-Iturbe, Evolution of dispersal in explicitly spatial metacommunities. J. Theor. Biol. **269**(1), 256–265 (2011)
25. L. Pechenik, H. Levine, Interfacial velocity corrections due to multiplicative noise. Phys. Rev. E **59**(4), 3893 (1999)
26. M. Plischke, B. Bergersen, *Equilibrium Statistical Physics* (World Scientific, Singapore, 2006)
27. J.B. Plotkin, M.D. Potts, N. Leslie, N. Manokaran, J. LaFrankie, P.S. Ashton, Species-area curves, spatial aggregation, and habitat specialization in tropical forests. J. Theor. Biol. **207**(1), 81–99 (2000)
28. J. Rosindell, S.J. Cornell, Species–area relationships from a spatially explicit neutral model in an infinite landscape. Ecol. Lett. **10**(7), 586–595 (2007)
29. J. Rosindell, S.J. Cornell, S.P. Hubbell, R.S. Etienne, Protracted speciation revitalizes the neutral theory of biodiversity. Ecol. Lett. **13**(6), 716–727 (2010)
30. J. Rosindell, S.P. Hubbell, R.S. Etienne, The unified neutral theory of biodiversity and biogeography at age ten. Trends Ecol. Evol. **26**(7), 340–348 (2011)
31. A.L. Šizling, D. Storch, Power-law species–area relationships and self-similar species distributions within finite areas. Ecol. Lett. **7**(1), 60–68 (2004)
32. D. Storch, P. Keil, W. Jetz, Universal species-area and endemics-area relationships at continental scales. Nature **488**(7409), 78–81 (2012)
33. N.G. Van Kampen, *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library, 3rd edn. (North Holland, Amsterdam, 2007)
34. I. Volkov, J.R. Banavar, S.P. Hubbell, A. Maritan, Neutral theory and relative species abundance in ecology. Nature **424**(6952), 1035–1037 (2003)
35. I. Volkov, J.R. Banavar, F. He, S. Hubbell, A. Maritan, Density dependence explains tree species abundance and diversity in tropical forests. Nature **438**(7068), 658–661 (2005)
36. I. Volkov, J.R. Banavar, S.P. Hubbell, A. Maritan, Patterns of relative species abundance in rainforests and coral reefs. Nature **450**(7166), 45–49 (2007)

Index

D. Holcman (ed.), *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, DOI 10.1007/978-3-319-62627-7