

Progress in the Chemistry of Organic Natural Products

A. Douglas Kinghorn · Heinz Falk
Simon Gibbons · Jun'ichi Kobayashi
Yoshinori Asakawa · Ji-Kai Liu *Editors*

110

Progress in the Chemistry of Organic Natural Products

Cheminformatics in Natural Product
Research

 Springer

Progress in the Chemistry of Organic Natural Products

Series Editors

A. Douglas Kinghorn, Columbus, OH, USA
Heinz Falk, Linz, Austria
Simon Gibbons, London, UK
Jun'ichi Kobayashi, Sapporo, Japan
Yoshinori Asakawa, Tokushima, Japan
Ji-Kai Liu, Wuhan, China

Advisory Editors

Giovanni Appendino, Novara, Italy
Roberto G. S. Berlinck, São Carlos, Brazil
Verena Dirsch, Wien, Austria
Agnieszka Ludwiczuk, Lublin, Poland
Rachel Mata, Mexico, Mexico
Nicholas H. Oberlies, Greensboro, USA
Deniz Tasdemir, Kiel, Germany
Dirk Trauner, New York, USA
Alvaro Viljoen, Pretoria, South Africa
Yang Ye, Shanghai, China

The volumes of this classic series, now referred to simply as “Zechmeister” after its founder, Laszlo Zechmeister, have appeared under the Springer Imprint ever since the series’ inauguration in 1938. It is therefore not really surprising to find out that the list of contributing authors, who were awarded a Nobel Prize, is quite long: Kurt Alder, Derek H.R. Barton, George Wells Beadle, Dorothy Crowfoot-Hodgkin, Otto Diels, Hans von Euler-Chelpin, Paul Karrer, Luis Federico Leloir, Linus Pauling, Vladimir Prelog, with Walter Norman Haworth and Adolf F.J. Butenandt serving as members of the editorial board.

The volumes contain contributions on various topics related to the origin, distribution, chemistry, synthesis, biochemistry, function or use of various classes of naturally occurring substances ranging from small molecules to biopolymers.

Each contribution is written by a recognized authority in the field and provides a comprehensive and up-to-date review of the topic in question. Addressed to biologists, technologists, and chemists alike, the series can be used by the expert as a source of information and literature citations and by the non-expert as a means of orientation in a rapidly developing discipline.

All contributions are listed in PubMed.

More information about this series at <http://www.springer.com/series/10169>

A. Douglas Kinghorn • Heinz Falk •
Simon Gibbons • Jun'ichi Kobayashi •
Yoshinori Asakawa • Ji-Kai Liu

Editors

Progress in the Chemistry of Organic Natural Products

Cheminformatics in Natural Product Research

Volume 110

With contributions by

F. D. Prieto-Martínez • U. Norinder • J. L. Medina-Franco

Y. Chen • C. de Bruyn Kops • J. Kirchmair

T. Rodrigues

T. Seidel • D. A. Schuetz • A. Garon • T. Langer

D. Reker


F. Mayr • C. Vieider • V. Temml • H. Stuppner • D. Schuster


B. Kirchweger • J. M. Rollinger




Springer

Editors


A. Douglas Kinghorn 
College of Pharmacy
The Ohio State University
Columbus, Ohio, USA

Heinz Falk 
Institute of Organic Chemistry
Johannes Kepler University
Linz, Austria

Simon Gibbons 
UCL School of Pharmacy
University College London, Research
London, United Kingdom

Jun'ichi Kobayashi
Grad. School of Pharmaceutical Science
Hokkaido University
Fukuoka, Japan

Yoshinori Asakawa 
Faculty of Pharmaceutical Sciences
Tokushima Bunri University
Tokushima, Japan

Ji-Kai Liu 
School of Pharmaceutical Sciences
South-Central Univ. for Nationalities
Wuhan, China

ISSN 2191-7043

ISSN 2192-4309 (electronic)

Progress in the Chemistry of Organic Natural Products

ISBN 978-3-030-14631-3

ISBN 978-3-030-14632-0 (eBook)

<https://doi.org/10.1007/978-3-030-14632-0>

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

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

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Foreword

“Big data” has emerged as one key term of the twenty-first century. Wikipedia, which itself is visible evidence of this development, defines the term as a “field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.”

It is therefore not surprising that also in the field of natural product chemistry over the last few decades, cheminformatic methods have evolved to analyze databases. The current volume of “Progress in the Chemistry of Organic Natural Products” presents a collection of contributions by authors who are experts in this field.

The first contribution (“Cheminformatics Explorations of Natural Products”) by José Medina-Franco and his colleagues from the National Autonomous University of Mexico gives a broad overview of cheminformatics strategies that may be used to mine natural product spaces for their potential biological activity, toxicity, or biodiversity.

The following chapter “Resources for Chemical, Biological, and Structural Data on Natural Products” is written by a young team working with Johannes Kirchmair from the University of Bergen (Norway) and University of Hamburg (Germany). Therein, they critically review approaches for using cheminformatic tools including virtual databases, physical natural product collections, and resources for biological and structural data on natural products.

The chapter “A Toolbox for the Identification of Modes of Action of Natural Products” by Tiago Rodrigues from the Instituto de Medicina Molecular João Lobo Antunes (Portugal) reviews cheminformatics tools for the identification of modes of action of natural products from molecular docking to machine-learning methods.

Thierry Langer and his team from the University of Vienna (Austria) provide a detailed introduction into pharmacophore-based techniques and the underlying concept that can be used in natural products chemistry and exemplify respective projects (“The Pharmacophore Concept and its Applications in Computer-Aided Drug Design”).

Daniel Reker from the Massachusetts Institute of Technology (USA) illuminates the relevance of natural fragments for drug discovery in his contribution “Cheminformatic Analysis of Natural Product Fragments.”

The chapter “Open Access Activity Prediction Tools for Natural Products. Case Study: hERG Blockers,” contributed by a team working with Daniela Schuster from the Paracelsus Medical University Salzburg and the University of Innsbruck (Austria), shows how potential toxicity caused by interference of natural products with the hERG potassium ion channel can be recognized by computational tools.

Finally, Benjamin Kirchweger and Judith Rollinger from the University of Vienna (Austria) analyze the strength, weaknesses, opportunities, and threats of cheminformatics methods that are used in natural product research (“A SWOT Analysis of Cheminformatics in Natural Product Research”).

In sum, Volume 110 offers a comprehensive and timely overview of how “big data” generated over the past decades in the form of natural product collections and databases can be mined by computational approaches to answer recurring issues. These include the molecular target identification of natural compounds as well as ligand identification for relevant macromolecular targets from the large pool of bioactive compounds from Nature, thus allowing us to assess their potential pharmacological and toxicological properties.

Vienna, Austria

Verena M. Dirsch

Contents

Cheminformatics Explorations of Natural Products	1
Fernando D. Prieto-Martínez, Ulf Norinder, and José L. Medina-Franco	
Resources for Chemical, Biological, and Structural Data on Natural Products	37
Ya Chen, Christina de Bruyn Kops, and Johannes Kirchmair	
A Toolbox for the Identification of Modes of Action of Natural Products	73
Tiago Rodrigues	
The Pharmacophore Concept and Its Applications in Computer-Aided Drug Design	99
Thomas Seidel, Doris A. Schuetz, Arthur Garon, and Thierry Langer	
Cheminformatic Analysis of Natural Product Fragments	143
Daniel Reker	
Open-Access Activity Prediction Tools for Natural Products. Case Study: hERG Blockers	177
Fabian Mayr, Christian Vieider, Veronika Temml, Hermann Stuppner, and Daniela Schuster	
A Strength-Weaknesses-Opportunities-Threats (SWOT) Analysis of Cheminformatics in Natural Product Research	239
Benjamin Kirchweiger and Judith M. Rollinger	

Cheminformatics Explorations of Natural Products



Fernando D. Prieto-Martínez, Ulf Norinder, and José L. Medina-Franco

Contents

1	Introduction	2
2	Mining Natural Product Spaces: Identification of Bioactive Compounds	4
2.1	Case Studies of Virtual Screening for Epigenetic Targets	7
2.1.1	Bromodomains	9
2.1.2	Sirtuins	11
2.1.3	DNA Methyltransferases	13
3	Toxicity Profile	15
3.1	Privileged or Promiscuous Natural Products?	17
3.2	Examples of Toxicity Profiling of Natural Product Databases	18
4	Diversity Analyses of Natural Products	19
4.1	Overview of Collections of Natural Products	19
4.2	Design of Nature-Inspired Compound Collections	20
4.3	Concept and Importance of Diversity Analysis	21
4.4	Representative Diversity Analysis of Natural Products	22
4.4.1	Global Analysis of Chemical Diversity	23
5	Conclusions and Future Directions	25
	References	26

F. D. Prieto-Martínez (✉) · J. L. Medina-Franco (✉)
Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico,
Mexico City, Mexico
e-mail: medinajl@unam.mx

U. Norinder
Department of Computer and Systems Sciences, Stockholm University, Kista, Sweden
Unit of Toxicology Sciences, Swetox, Karolinska Institutet, Södertälje, Sweden
e-mail: ulfn@dsv.su.se

© Springer Nature Switzerland AG 2019

A. D. Kinghorn, H. Falk, S. Gibbons, J. Kobayashi, Y. Asakawa, J.-K. Liu (eds.),
Progress in the Chemistry of Organic Natural Products, Vol. 110,
https://doi.org/10.1007/978-3-030-14632-0_1

Abbreviations

BRD	Bromodomain
CDPs	Consensus Diversity Plots
DNMT	DNA methyltransferase
FDA	Food and Drug Administration
HDAC	Histone deacetylase
hERG	Human ether-a-go-go-related gene ion-channel
IMPS	Invalid metabolic panaceas
MACCS	Molecular Access System
PAINS	Pan-Assay Interference compounds
PCA	Principal component analysis
SAH	S-adenosyl homocysteine
SAM	S-adenosyl methionine
SMILES	Simplified Molecular Input Line Entries
TCM	Traditional Chinese Medicine
UNPD	Universal Natural Products Database

1 Introduction

Natural products have intimate relationships with medicine and chemistry, with various examples from ancient civilizations throughout history. Most of these uses include those in traditional or herbal medicine, to which also mystical properties to the plants or fungi concerned have sometimes been attributed. For example, sage is a herb that was thought to ward off evil. Nowadays, it is known that sage possesses several biological effects, for example, antibacterial, antioxidant, and cholinergic [1]. In a similar manner, other traditional uses have been validated by scientific research [2–5].

As such, natural sources have driven the early stages of medicinal chemistry and drug discovery, yielding valuable therapeutic agents still in use today. Prominent examples of drugs approved for clinical use from natural sources include, but are not limited to, penicillin, pilocarpine, reserpine, and salicylic acid. Furthermore, the role of natural products as novel avenues for therapy increased after the so-called Golden Age of Antibiotics (circa 1960) when the larger companies in the pharmaceutical industry began the development of numerous projects, searching for molecules with diverse bioactivities [6]. However, the “golden age” of natural products as antibiotics was quite short, since most companies reduced such endeavors by the turn of the twenty-first century [7]. Several reasons have been given that help explain the decreased enthusiasm of pharmaceutical companies to work on natural products. Two major points are the inherent complexity of crude extract compound mixtures and the slowness of natural product optimization [8]. Additionally, with the rapid development of combinatorial chemistry and high-throughput methods, the search

for chemical diversity was considered a solved problem. Unfortunately, this has not been the case, as it has been shown that combinatorial collections tend to get trapped in the same area of chemical space [9]. Moreover, even with the ability to produce compounds in high numbers, only a handful of Food and Drug Administration (FDA)-approved drugs come from such methods [10]. Therefore, it can be argued that the solution of the problem “quantity over quality” is “quality over quantity”.

As a result, natural products have seen a “rebirth” with novel methods and synthesis strategies to produce diverse collections [11]. Additionally, in most cases, vegetal sources are the major players in natural product research. Thus, other sources like marine, bacterial, and fungal metabolites offer untapped potential [12, 13]. As recently reviewed, there are several recently approved drugs that are natural products or are synthetic analogs of hit compounds initially identified from natural sources. A clear and recent example is the fungal metabolite migalastat (Galafold[®]) approved in 2018 for the treatment of Fabry disease [14].

Due to these considerations, current efforts involve multidisciplinary approaches, which help mitigate the problems inherent to natural products. This mainly focuses on the improvement of extraction, isolation, and quality control of metabolites, including “omics technology” [15]. Nonetheless, other technological approaches have arisen. Take, for example, the high volume of information available on natural products and their activities. We now live in an era of “big data”, with different dedicated repositories [16]. The rational and effective mining of such databases could yield important breakthroughs.

It is well known that many natural products exert multiple effects *in vitro*, and, because of this promiscuous nature, some classes of natural products are among the Pan Assay Interference Compounds (PAINS, see Sect. 3) [17]. It follows that a screening campaign might well filter scaffolds of natural products to identify promising ones, while also discarding PAIN-like moieties. In practice, this can be accomplished rather easily, by conducting a virtual screening that is an *in silico* method (part of cheminformatics) aimed at selecting compounds with potential biological activity.

A rather “young” discipline, cheminformatics, is envisioned as the answer for chemical information problems using several numerical, statistical, and physico-chemical methods to work with two- and three-dimensional chemical structures [18]. This aims to optimize resources more effectively and to focus on the more viable molecules. Therefore, cheminformatics relies heavily on concepts like chemical space, molecular similarity, and chemical representation [19]. More recently, the scope of cheminformatics has shifted toward *in silico* evaluation, using molecular modeling approaches and machine learning.

The goal of this chapter is to discuss the progress of selected cheminformatic strategies to further advance the identification of bioactive molecules from natural origin. This contribution is organized in five major sections. After this introduction, Sect. 2 discusses examples of mining the space of natural products using several virtual screening strategies, including similarity searching, automated docking, and consensus methods. In this section, case studies are described of virtual screening for the identification of bioactive molecules against epigenetic targets. Section 3

discusses the *in silico* toxicity profiling of natural product datasets. Next, Sect. 4 covers the analysis of the chemical diversity and coverage in chemical space as well as the design of natural product-like molecules and natural product mimetics. Section 5 presents summary conclusions and perspectives.

2 Mining Natural Product Spaces: Identification of Bioactive Compounds

As stated, virtual screening aims to evaluate the potential of a molecule as a biological agent. This can be achieved in several ways; some of these are listed in Table 1.

Usually, a virtual screening protocol involves various methods in consecutive order, trying to filter large databases to “cherry-pick” putative ligands of interest. Thus far, virtual screening has been applied successfully to identify hit compounds that are usually later optimized [26–28].

In the early days of *in silico* research, the quintessential approaches were descriptor-based, mostly inspired by the success of the Hansch-Fujita method. This led to the birth of Quantitative Structure Activity Relationships (QSAR) and their more refined counterparts: CoMFA and CoMSIA [29]. A prominent success

Table 1 Representative computational methods and concepts used for virtual screening

Method/concept	Brief description	Refs.
Chemical space	Abstract representation of compounds, using different descriptors. This allows the profiling of chemical collections	[20]
Molecular similarity	Using graph decomposition, molecular structures are codified as vectors. These in turn can be compared using different equations to measure similarity	[21]
QSAR	Mathematical models supported by descriptors that quantify the impact of substituents in biological activity. Their main aim is the prediction of biological activity	[22]
Molecular docking	Simulation that approximates protein-ligand binding. This is accomplished by the conformational searches of ligands and the evaluation of these using dG values as criteria	[23]
Molecular dynamics	Physical simulations that allow the study of protein behavior, using equations of motion and potential energy functions (forcefields)	[24]
Free energy perturbations	Derivatives of molecular dynamics, in this case the simulation goes across a thermodynamic cycle. This can be used for the approximation of binding energy and the change in its value due to fragment changes	[25]

case being the Lipinski Rule of Five, which describes a general profile of “drug-like” molecules with optimal bioavailability (no more than 5 hydrogen bond donors, no more than 10 hydrogen bond acceptors, $M \leq 500$, $\log P \leq 5$) [30]. Alas, it can be argued that over-reliance on such approaches has led to molecular attrition [31]. In addition, it has been shown that the overall performance of descriptor-based classification depends on the correct assessment of relevant properties [32].

On the other hand, there are receptor-based approaches, with the most well-known of them being molecular docking. One such technique uses the GRID method, developed by Goodford et al., which generates molecular interaction maps in protein cavities [33]. Hence, docking can be used to model drug–protein complexes and perhaps the most appealing aspect of this, the calculation of relative binding energies.

Even so, molecular docking has critical points that may be often overlooked by naive users, for example, structure selection, protein preparation, the inclusion of water molecules and metal ions, and protein flexibility [23, 34]. Furthermore, one of the most important flaws in molecular docking is the pose versus scoring phenomena that are related to the uncertainty of significant results without the proper knowledge of the binding site. Consequently, some protocols and good practices have been proposed for reliable results [35, 36]. In this sense, proper ligand selection has been suggested as a preferred method for docking candidate selection [37].

Of the several approaches for molecule mining, chemical similarity is perhaps the most powerful. Most chemists have encountered this principle, sometimes inadvertently. The rather simple axiom, “similar structures share similar activities,” holds significantly true in a pharmacological context. In practice, chemical similarity provides a tool for systematic and objective comparison of compound pairs. To do this, chemical structures are codified as strings, known as Simplified Molecular Input Line Entries (SMILES). Then follows a comparison based on topology or fragment substructures, commonly performed with the Tanimoto coefficient to compute similarity values [38].

Without doubt, similarity methods have improved the overall capacities of virtual screening, with recent examples of success in the literature [39]. Nevertheless, molecular similarity is not fail-proof due to structure–activity relationship heterogeneity. More explicitly, this refers to the existence of activity-cliffs, that is, molecules with a known active scaffold that loses its effect with small modifications (pyridine instead of benzene ring) as with compounds **1a** and **1b** shown in Fig. 1.

This phenomenon deeply impacts the performance of virtual screening as a whole, not just similarity methods [40]. Accordingly, the best results of virtual screening campaigns are obtained by complementary approaches, also known as consensus [41].

Virtual screening protocols may be implemented rather easily and with such potential, they have been adopted in natural product research. Correspondingly, screening and optimization of natural products has benefited from computational tools. In turn, computational chemists saw the potential of natural products as privileged scaffolds for lead searching, ending in a symbiotic relationship early on. As may be expected, there have been some inherent difficulties and successes

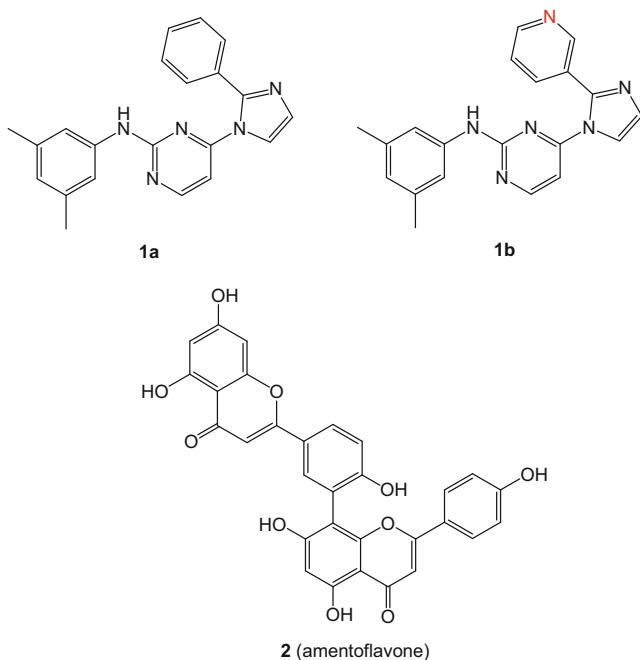


Fig. 1 Example of an activity cliff, with the most potent structure being **1b**. In this case, the difference in activity between **1a** and **1b** is almost 400 times. Of note, this large change in activity is due to a single heteroatom. Below, structural formula of amentoflavone (**2**)

along the way. Still, this interdisciplinary environment has led to the development of public repositories and the overall improvement of computational algorithms [42].

Generally, the proposal or study of putative mechanisms of action is the main goal of computational methods in natural product research. For example, DNA topoisomerases have been studied with a wide array of natural products, identifying interaction patterns crucial to enzyme inhibition [43]. These concepts have been scaled further as “target fishing” or reverse virtual screening. In this case, the molecule of interest is used as filter, that is, it is evaluated against several targets to identify significant activities. The value of such studies cannot be overstated, as their utility may range from structure–activity relationship optimization to multi-activity map pathways [44].

Likewise, molecular modeling tools have been used to identify natural product leads with micromolar activities in targets such as acetylcholinesterase (AChE), cytochrome P-450, angiotensin-converting enzyme 2 (ACE-2), kinase CK2, and estrogen receptor- β [42]. On the other hand, consensus protocols have been successful in the screening of marine compounds with assorted activities [13].

As may be seen, natural product mining with virtual screening protocols has proven effective. Of course, there are more examples in different fields, but we

consider that among them, the epigenome provides an interesting application for natural products as chemoprotective agents. Here, we discuss recent applications with emphasis on epigenetic targets that are emerging as promising targets for the treatment of several diseases [45–49].

2.1 Case Studies of Virtual Screening for Epigenetic Targets

Epigenetics has become an attractive area of study, first described in 1940 by Conrad Waddington [50]. It refers to heritable changes in gene expression that occur independent of alterations in DNA sequence, but are rather based on modifications of histone proteins or nucleic acids. Since its description, epigenetics is linked to factors such as diet or the environment to explain the biogenesis of some diseases [51].

Currently, epigenetics has provided a novel approach to search for therapies in the treatment of cancer, diabetes, hypertension, or even Alzheimer’s disease. Still, epigenetic modulation is not “black or white”, as several epigenetically modifying enzymes modulate a wide array of physiological functions. In addition, the epi-pocketome continues to grow at steady pace, increasing target diversity and complexity [52, 53]. Hence, the overall safety and scope of epi-therapies are yet quite blurry [54].

Consequently, the search for epi-modulators is not limited to drugs but is focused on the identification of probes [55, 56]. In this context, natural products have taken a prominent role in the field, serving as leads or even templates to understand epi-pharmacology. Some examples (3–11) of epi-modulators are presented in Fig. 2.

Of note, flavonoids have a privileged place among natural products as therapeutic agents. Often regarded as natural polydrugs, this scaffold has a plethora of biologic actions beyond their antioxidant potential [57]. Considering their abundance in human diet, flavonoids have a well-documented nutraceutical potential [58].

In the next sub-sections, we further comment on some case studies where natural products are involved in serving as leads or to uncover interesting structure–activity relationships.

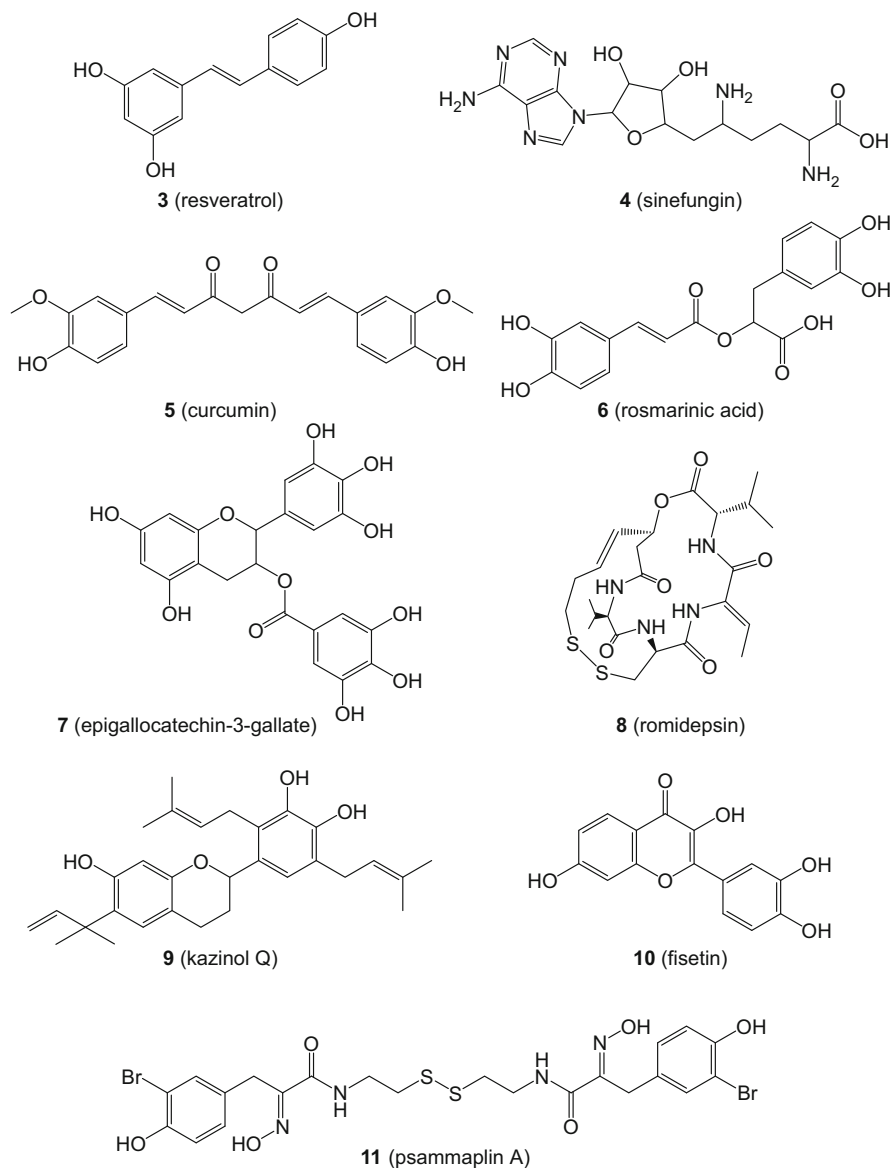


Fig. 2 Illustrative examples of natural products reported as epigenetic modulators, as identified by direct or indirect mechanisms. Most of the examples have supportive *in silico* modeling studies that help to explain their effect

2.1.1 Bromodomains

Bromodomains (BRDs) are small proteins (around 120 residues) that are classified as epi-readers, that is, enzymes for which the function is focused on recognizing patterns of a given moiety. In this case, bromodomains identify acetylated lysine residues [59]. Currently, over 60 isoforms of bromodomains have been identified from the human proteome; of those, bromodomain and extraterminal domains (BETs) have attracted the most interest so far. This is mainly due to their relation to cancer cell lines and inflammatory processes [60].

One of the pitfalls in bromodomain inhibition is the lack of structural diversity in current inhibitors [61]. As a result of this, there is an ongoing search for novel inhibitors of these targets. Additionally, BET isoforms exhibit high values of sequence similarity in their binding site, making the search more difficult for selective and potent inhibitors.

Recent endeavors in the field include fragment-based virtual screening [62], lead optimization based on receptor structure [63], development of bivalent inhibitors [64], and molecular dynamics of active sites [65]. With this background, our group focused on molecular modeling methods to further advance the understanding of BET inhibition [66].

Following a virtual screening protocol using molecular similarity and docking, two hits were identified. The more promising was amentoflavone (**2**) (Fig. 1), a biflavonoid produced by *Ginkgo biloba* and *Hypericum perforatum* among other plants, with previous reports of antitumor-related activity [67, 68]. Similarly, other groups identified the flavonoid scaffold as a putative ligand for bromodomains [69, 70]. Yet, this was the first report for biflavonoids, which is interesting due to their atropisomeric properties [71]. In addition, all these studies suggested that flavonoids bind at the ZA channel (a flexible region connecting the Z and A loops). This region has been suggested as significant for selectivity due to its interaction with a conserved water network [72].

Further characterization was performed with molecular dynamics simulations, which showed that amentoflavone (**2**) can interact with D145, a residue specific to BRD4-BD1 [73]. This is an interesting observation considering that RVX-297 (a quinazoline) is a specific inhibitor of BRD4-BD2 [74]. Biological evaluation of amentoflavone showed an IC_{50} in the micromolar range, with evidence suggesting selectivity for BRD4-BD1 [75].

Thus, it can be stated that atropisomerism provides positive contacts for BRD4-specific inhibition. As a proof of concept, Fig. 3 presents protein–ligand interactions with selected biflavonoids obtained by molecular dynamics. This shows that indeed, the spatial arrangement and conformational freedom of ligands favor their interaction to D145.

Recently, isothermal titration calorimetry assays have shown that binding in the pocket of BETs is mostly enthalpy driven [76]. This in addition to the flexibility of the ZA channel suggests that constrained structures can show BET selectivity and specificity. This is a notable observation considering the rather “simple” scaffold of

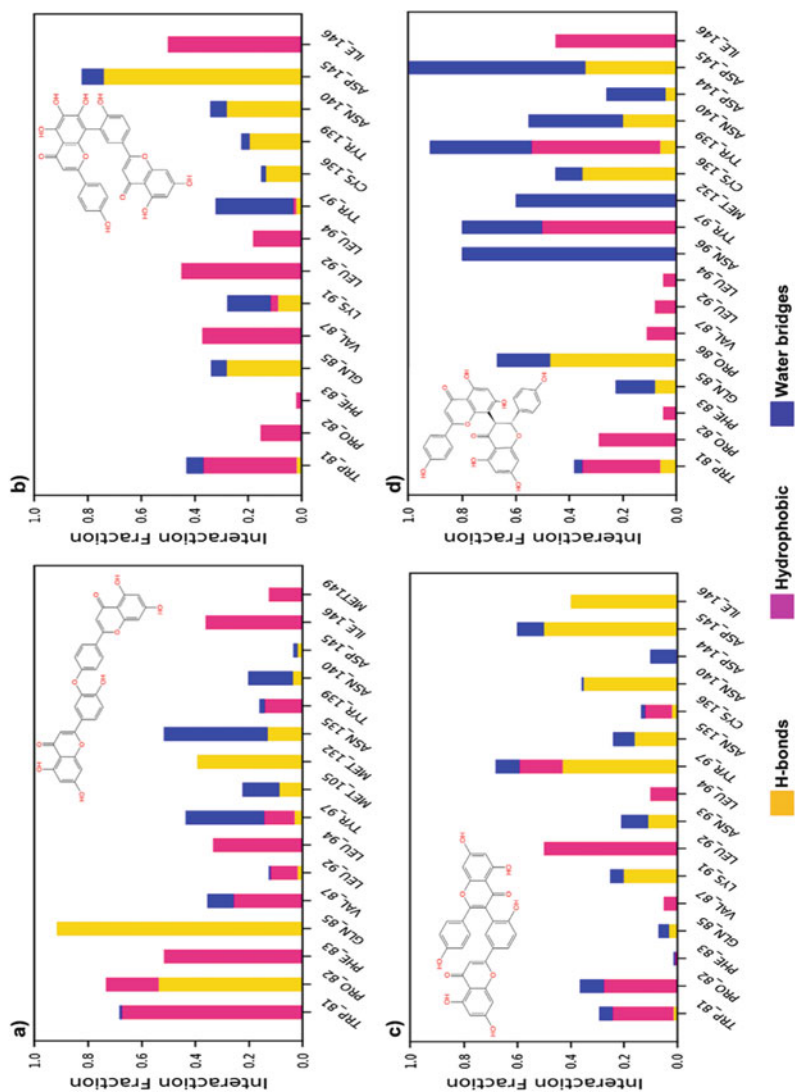


Fig. 3 Protein-ligand interactions as obtained from molecular dynamics of BRD4-biflavonoid complexes. (a) BRD4-ochraflavone, (b) BRD4-taiwanoflavone, (c) BRD4-sumaflavone, (d) BRD4-talbotaflavone

flavones. Nevertheless, this shows the undeniable potential of natural products, not just as leads but as pharmacophore templates.

2.1.2 Sirtuins

While not yet discussed in the previous Section on bromodomains, histone acetylation is crucial for chromatin opening. This happens as a result of the recruitment of histone acetyl transferases, and to reverse this process, histone deacetylases (HDACs). The latter are intensively studied to develop novel therapies for several cancer lines, by reactivating silenced genes [77]. Currently, 18 HDAC isoforms are classified into four different classes in regard to their homology to yeast proteins. Class III is the only one for which the function relies on nicotinic adenine dinucleotide (NAD⁺), also known as sirtuins due to their relation to Sir2 [78].

There are seven isoforms of sirtuins in humans expressed at different cellular locations, with highly conserved active sites, but functionally different structures and domains [79]. Recently, it has been shown that sirtuins exert functions beyond epigenetic silencing [80]. For example, sirtuins have an active role in DNA protection and repair by several mechanisms, which include PARP activation, glutamine anaplerosis, reactive oxygen species, and activation of reactive oxygen species neutralizing enzymes [81]. Moreover, sirtuin expression has a direct correlation with caloric restriction. This has been related to extended life span and overall health status provided by NAD⁺ upregulation [82]. Hence, the investigation of sirtuins becomes quite interesting, as the focus diverges for the search of both inhibitors and activators, according to the effect desired.

One of the first inhibitors of the HDACs was romidepsin (**8**), a depsipeptide with a disulfide bond and a caged structure, identified from *Chromobacterium violaceum* [83]. In subsequent studies, it was shown that romidepsin activity was mediated by rupture of the disulfide bond, followed by covalent inhibition of catalytic zinc ions [84]. As a result of this, **8** has pleiotropic effects via pan-HDAC inhibition [85]. Romidepsin (**8**) has been approved by the FDA for the treatment of T-cell lymphoma [86].

Psammaphin A (**11**) also contains a disulfide bond, which gives it a potent but nonspecific inhibition of HDACs. Synthesis optimization of this structure led to UVI5008, a compound with the added capacity to inhibit SIRT1/2 [87].

As such, with the off-target effects and nonspecific binding, some researchers have used in silico methods in order to further investigate the inhibition of sirtuins. Early studies focused on splitomicin, an inhibitor of yeast sirtuins. Using molecular docking and molecular mechanics methods, structure–activity relationships were obtained for splitomicin derivatives. These studies provided insight into the rationale behind the activity of (*R*)-enantiomers of these scaffolds, which were also non-competitive SIRT2 inhibitors [88].

Kokkonen et al. [89] conducted a 3D QSAR study based on SIRT1. Using the CoMFA method a model of significant predictive power was obtained, which resulted in peptide-like ligands for SIRT1 with *IC*₅₀ values around 10 μM. Following

a subsequent ligand-based virtual screening by Sun et al. [90] using data from public repositories and literature records, 36 representative ligands were selected to obtain binding models using molecular docking. With this model, 12 compounds from Traditional Chinese Medicine were identified as putative ligands of SIRT1. That same year a classic screening of the same database was carried out, identifying four actives out of 19 candidates for SIRT1 activation [91].

A recent study by Karam et al. [92] presented a virtual screening protocol followed by in vitro testing, with a focus on SIRT1, 2, and 3. Using a dataset of African-derived natural products (p-ANAPL), 13 compounds were selected by molecular docking. Seven of these compounds contained a chalcone scaffold with modest activity against SIRT1 and 2. Further modeling showed that the putative binding poses correlate with known crystallographic structures.

Another isoform of interest is SIRT6, as it is related to inflammatory and aging processes. Several studies in mice have shown the importance of this enzyme, particularly its role in cardioprotective mechanisms [93]. Rahmasto-Rilla et al. [94] focused on several flavonoids as putative SIRT6 modulators. The authors of this work used first in vitro screening to identify inhibition/activation of this enzyme. Remarkably, the nature of the modulation was concentration-dependent, with anthocyanidins being identified as effective activators of SIRT6. To gain further insights, molecular docking and in silico residue mutations were carried out, identifying the putative site for activators and the possible mechanism being conformational changes induced by the amino acid residues G156, D185, W186, E187, and D188.

Finally, we discuss the role of sirtuin inhibitors as putative antiparasitic agents. This arises from the phylogenetic characterization of sirtuins, identifying SIR2 homologous enzymes in pathogens, for example, *Toxoplasma* spp., *Plasmodium* spp., *Trypanosoma cruzi*, *Leishmania* spp., and *Trichomonas vaginalis* [95]. This opens an avenue for novel therapies of the so-called neglected diseases, as it has been shown that these enzymes have direct relationship with growth and infectivity of pathogens [96, 97].

In this regard, in silico modeling has been used to assess the viability of these macromolecules as potential targets for the treatment of infections. Mostly by homology modeling, studies have suggested that parasitic sirtuins have enough differences from human isoforms to warrant low toxicity [98, 99].

With this in mind, and as a proof of concept, we selected *Trypanosoma cruzi* Sir2-related protein 3 (TcSir2rp3), as a potential target for the treatment of Chagas disease, and conducted representative virtual screening. Beginning with a homology model for *T. cruzi*, sirtuin coupled with NAD⁺, to conduct molecular docking with putative ligands. Also, we focused on flavonoids, due to their background discussed above.

2.1.3 DNA Methyltransferases

Deoxyribonucleic acid may be modified by the addition of methyl groups. This may be conducted over the CpG islands, specifically position 5 of cytosine nucleotides. These regions on DNA are related to gene promoters, so methylation-induced silencing is a recurring feature in most types of cancer [100]. This process involves de novo methylation carried out by the enzymes DNA methyltransferases (DNMTs) 3A and DNMT3B, while “maintenance” is done by the isoform DNMT1. Abnormal function of DNMTs has been related to other malignancies, such as asthma, lupus erythematosus, and myelodysplastic syndrome [101].

An indirect inhibition of DNA methylation, with the use of the nucleotide 5-azacytidine, resulted in re-expression of silenced genes and inhibition of tumor growth [49]. As a result of this, analogs of *S*-adenosyl methionine and *S*-adenosyl homocysteine (SAM/SAH, respectively) have been studied to uncover the mechanisms of methyltransferases [102]. Sinefungin, a natural analogue of SAM is a pan-inhibitor of methyltransferases that continues to serve as template for rational design due to the “transition state model” presented earlier [103].

Nevertheless, nucleotide derivatives possess poor bioavailability and high toxicity, which necessitated research for non-nucleotide scaffolds [104]. Following the example of sinefungin, other natural products have been studied as direct or indirect demethylating agents. Phenolic compounds have a prominent place in these endeavors, as various studies have shown strong evidence of the chemoprotective role of these dietary compounds. Examples include (Figs. 2 and 4): genistein (**15**), rosmarinic acid (**6**), baicalein (**20**), and galangin (**21**); most of them exert indirect inhibition of DNMT1 by SAH accumulation [105]. Among these compounds, resveratrol (**3**) stands out, posing multi-target activities. A recent study by Maugeri et al. provided evidence of resveratrol modulation of SIRT1 and DNMT [106]. This serves as further evidence of the potential of **3** beyond its antioxidant capacities.

Using (*E*)-resveratrol analogs, the study of Aldawsari et al. showed that salicylate moieties provide putative DNMT3 selectivity [107]. By means of molecular modeling and in vitro testing it was assessed that these analogues may have activity independent of SAH, with an increased potency when compared to the parent compound.

Similarly, kazinol Q (**9**), a hydroxy-chromane derivative, showed antiproliferative activity at 10 μ M. Using molecular docking, it was shown that **9** binds to DNMT1 at the SAM site, sharing pharmacophoric traits with epigallocatechin-3-gallate (EGCG), despite the lack of a galloyl moiety [108].

As demonstrated above, natural products continue to offer numerous leads for epigenetic modulation. A focus toward multi-target activity and interdisciplinary research should together continue to uncover other mechanisms such as protein-protein interaction (PPI) modulation. However, the possible toxicity of natural products may still be an issue, as it is a main problem in drug discovery. Hence, in the next section, we address some of the advances and challenges to predict toxicity.

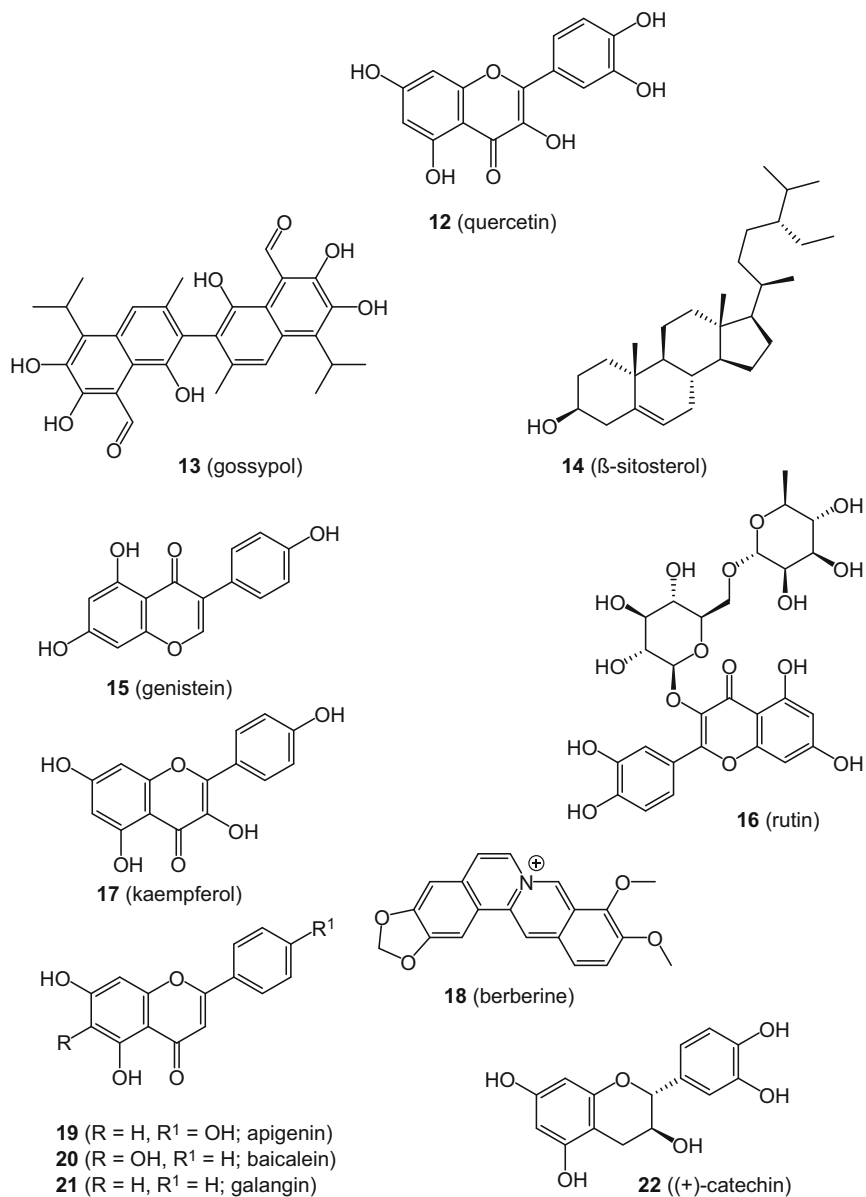


Fig. 4 Chemical structures of ten invalid metabolic panaceas (IMPs), a category that also includes curcumin (5)

3 Toxicity Profile

Despite the fact that natural products are regarded by the public domain as “safe” because they are “natural compounds” and indeed have been strongly associated with many health benefits, they can contain undesirable, for example, reactive or functional groups. They may also have other toxicological and other properties rendering them not suitable for drug discovery or human consumption such as preservatives or flavoring compounds. Certainly, there are secondary metabolites that are used as pesticides and are toxic.

In drug discovery, calculating or whenever feasible measuring or quantifying experimentally the toxicity profile of chemical compounds is mandatory. In the early stages of drug development, it is common to assess the toxicity related to cytochrome P450 or the human ether-a-go-go-related gene ion-channel (hERG). In later stages, other toxicity endpoints are commonly evaluated such as skin sensitization, potential for genotoxicity and carcinogenicity [109, 110]. This is because many research programs have failed due to toxicity concerns [110]. One of the strategies in order to anticipate toxicity issues is applying commercial, public or in-house algorithms [111, 112]. Indeed, the serious toxicity issues in drug discovery have boosted the need to develop tools to reliably and rapidly predict toxicity endpoints of compounds. Despite the fact that much progress has been made in *in silico* toxicology, this research area is still under development [110]. In this regard, it is relevant to bear in mind that accurate models become more challenging to develop as the complexity of the toxicity endpoint increases. Complex endpoints are characterized by having various mechanisms of action, that is, due to the interaction of one compound with multiple targets (“polypharmacology”) [113] or the interaction of multiple ligands with the same target (“polyspecificity”) [114], or the combination of both such as the case for certain fragrances (Hernández-Alvarado RB et al. 2019, personal communication). Moreover, the biggest challenge in toxicity modeling is that all chemical compounds are toxic at some level. Therefore, it is expected that a computational approach would be able to predict the type and level of toxicity. As commented by Gleeson et al., the prediction of the absolute toxic potential of a compound, either from *in silico* or animal models, is very difficult because there are a large number of ways in which toxicity (related to the primary pharmacology or many secondary pathways) can arise [110].

For practical purposes in many current drug discovery projects, structural alerts are used to rapidly identify small molecules that are reactive under common test conditions [115] or are associated with other undesirable properties [116]. These types of compounds have been termed PAINS in the literature (see above). The importance of PAINS structural alerts in natural product research for drug discovery has been discussed extensively by Baell [117].

In this context, it is essential to study and distinguish the concentration and the mechanism of toxicity of natural products. There are several studies that have been published with the aim of estimating the toxicity profile of natural product datasets. Table 2 summarizes representative work of *in silico* profiling of natural products and

Table 2 Examples of recent cheminformatic toxicity-related analysis of datasets of natural products

Study	Outcome	Refs.
In silico toxicological screening of natural products	This study compares the predicted vs. experimental toxicity profile for the naturally occurring dietary chemicals: estragole, pulegone, aristolochic acid I, lipoic acid, 1-octacosanol, and epicatechin. It was found that consensus predictions appear to be more accurate than the use of only one or two software programs. In silico results were in agreement with the experimental toxicity data	[118]
In silico toxicity profiling of natural product compound libraries from African flora	Analysis of the diversity and chemical toxicity assessment of three chemical collections of compounds from African flora. The predictions were done through the identification of chemical structural alerts. It was concluded that only a small fraction of the libraries could have toxicities beyond acceptable limits	[119]
In silico prediction of the toxic potential of lupeol	Lupeol is a triterpenoid found in many plant species. The interaction of lupeol and 11 of its analogues toward a series of 16 proteins known or suspected to trigger adverse effects was investigated. It was found that there is a moderate toxic potential for lupeol and some of its analogues, by targeting and binding to nuclear receptors involved in fertility	[120]
Toxicity assessment of natural products from Mexican plants with antinociceptive activity	Assessment of the toxicological profile of molecules with analgesic activity from the UNIIQUIM database. Most of the compounds are likely to interact with opioid receptors. The predicted acute toxicity is low and none is predicted as mutagenic	[121]
PAINS alerts of a Brazilian dataset and other reference datasets	A large number of molecules in NuBBE _{DB} are promising sources of molecules for medicinal chemistry and drug discovery projects	[122]
Promiscuity predictions for 208,000 natural products	Predictions of promiscuous compounds with the free online server Hit Dexter 2.0. Overall, flavonoids, in particular chalcones, are predicted as highly promiscuous. In contrast, alkaloids are predicted to be less promiscuous in general	[116]

computer-aided prediction of their toxicity profile. A representative study is further discussed below.

A visual representation of 24 ADME (absorption, distribution, metabolism, and elimination)-related properties for a TCM database [123] and natural products from the ZINC database [124] was obtained with principal component analysis (PCA). The so-called ADME space of the natural product collections was compared to a collection of approved drugs, commercial vendor compounds, a general diverse collection obtained from the National Cancer Institute database, and combinatorial collections. It was concluded that TCM covers a vast region of this property space,

including areas uncharted by drugs. Natural products from ZINC occupy the same area as drugs [123].

Physicochemical properties along with sub-structural features, for example, functional groups are also used as criteria to filter out compounds with potential toxicity issues early in the drug discovery process. To exemplify this point in recent work, Saldívar-González et al. classified seven natural product collections into six subsets including drug-like, extended drug-like, fragment-like, lead-like, PPI-like, and PAINS [122]. The collections were 2214 compounds from Brazil assembled in the NuBBE database, that is, the first collections of natural products of Brazilian biodiversity, with 473 cyanobacteria and 206 fungal metabolites, 6253 marine natural products, 4103 purified natural product screening compounds, 26,318 semi-synthetic molecules (the last two are commercially available for screening), 17,986 compounds from TCM, and 209,574 molecules in the Universal Natural Products Database (UNPD). Overall, it was found that all seven natural product types had a similar profile except cyanobacteria metabolites. In particular, it was concluded that the NuBBE database had a small percentage of PAINS molecules. In turn, cyanobacteria metabolites had a small fraction of drug-, extended drug-, and lead-like molecules with an increased fraction of PPI-like compounds.

Furthermore, in a recent investigation, Storck et al. profiled approximately 208,000 natural products with a new generation of machine-learning models to identify frequent hitters. The models are freely accessible through the web service Hit Dexter 2.0 [116]. Among the different results, it was found that there was a large percentage of flavonoids (more than 60% of the compounds analyzed) that were found to be promiscuous and approximately 20% highly promiscuous. Of the different flavonoids, chalcones showed the highest rates of promiscuity. In contrast to the predictions for flavonoids, the predictions found by Hit Dexter 2.0 suggested that alkaloids were much less promiscuous [116].

3.1 *Privileged or Promiscuous Natural Products?*

For some natural products, there is a debate and fine line between highly active or privileged compounds with numerous associated health-related benefits or non-specificity (or high reactivity) [125]. Perhaps one of the most notorious examples in this regard is curcumin (**5**), a constituent of turmeric (*Curcuma longa*), a traditional medicine. Curcumin (**5**) has been classified as both a PAIN [117] and “invalid metabolic panacea” (IMP) compound [126]. Despite the fact there are a large number of reports associating **5** with a plethora of biological activities, there are no conclusive positive results in randomized, placebo-controlled clinical trials for any studied indication as recently discussed by Nelson et al. [127]. Figure 4 shows the chemical structures of nine additional natural products regarded as IMPs in the study by Bisson et al. [126], namely: quercetin (**12**); gossypol (**13**); β -sitosterol (**14**); genistein (**15**); rutin (**16**); kaempferol (**17**); berberine (**18**); apigenin (**19**); and (+)-catechin (**22**) (selected from a list of 39 compounds in total).

3.2 *Examples of Toxicity Profiling of Natural Product Databases*

As commented above, it is common to evaluate the toxicity related to hERG during the first steps of drug development. Inhibition of this ion channel has been associated with a potentially fatal cardiac arrhythmia, Torsades de Pointes [128]. Several varied experimental tests are routinely used to evaluate hERG inhibitory potential. A number of in silico methods have been developed to assess hHERG inhibition as reviewed by Gleeson et al. [110]. In turn, the *Salmonella*/microsome assay (Ames assay) is a bacterial short-term test for identification of carcinogens using mutagenicity in bacteria as an endpoint. It is one of the most widely used short-term tests. A high (but not conclusive) association has been found between carcinogenicity in animals and mutagenicity in the Ames assay. Despite the fact there is still controversy over the value of *Salmonella*/microsome assay results in risk assessment, the results of the Ames assay can provide valuable information to aid in the development of further studies, and may form part of the data, which can be used in evaluating potential biological effects or projected lack of adverse effects [129].

To further illustrate the toxicity profile of natural product datasets of general interest, Table 3 summarizes the predicted Ames' toxicity and hERG affinity of six datasets of natural products previously profiled in terms of structural and whole-molecule properties (vide supra, [14]). As reference, the calculations were done for a dataset of 1806 drugs approved for clinical use. The curation of the datasets is described in detail by González-Saldívar et al. [122]. These calculations were done using in-house algorithms and the analysis revealed that the cyanobacteria metabolites contained a small fraction of compounds with predicted Ames mutagenicity (2.3%) followed by compounds in the semi-synthetic collection NATx (3.3%). The two datasets with the largest fraction of compounds with calculated Ames mutagenicity were NuBBE database and fungal metabolites (10.4 and 10.7%, respectively) which represent in each case a higher proportion than the approved drugs for clinical use also investigated (8.6%).

Regarding the predicted toxicity due to hERG affinity, all six natural product datasets had lower proportions of compounds predicted with high affinity as compared to approved drugs (13.5%). In particular, the datasets with the lowest proportion were fungal metabolites (0.5%) followed by marine and natural products from the commercial screening collection MEGX (1.2 and 1.3%). These results further support that, overall, the six natural product collections can be used as a starting point in drug discovery studies, for instance, in virtual screening to identify potential hits. Of course, the prediction of the toxicity (such as illustrated in Table 3) can be used as a guide to filter compounds for selection.

Table 3 Examples of in silico Ames toxicity and hHERG affinity profiles of six natural product datasets and compared to drugs approved for clinical use

Ames							
Dataset	Size	Yes	Yes (%)	No	No (%)	NA	NA (%)
Cyanobacteria	473	11	2.3	456	96.4	6	1.3
Fungi	206	22	10.7	180	87.4	4	1.9
MEG x	4103	333	8.1	3660	89.2	110	2.7
NAT x	26,318	860	3.3	25071	95.3	388	1.5
NuBBE	2214	231	10.4	1925	86.9	58	2.6
Marine	6253	420	6.7	5700	91.2	133	2.1
Approved drugs	1806	156	8.6	1610	89.1	39	2.2
hHERG ^a							
Dataset	Size	Yes	Yes (%)	No	No (%)	Inconclusive	NA (%)
Cyanobacteria	473	8	1.7	445	94.1	20	4.2
Fungi	206	1	0.5	202	98.1	3	1.5
MEG x	4103	53	1.3	3977	96.9	73	1.8
NAT x	26,318	2841	10.8	21,008	79.8	2469	9.4
NuBBE	2214	44	2.0	2054	92.8	116	5.2
Marine	6253	73	1.2	5924	94.7	256	4.1
Approved drugs	1806	243	13.5	1435	79.5	126 (+2 empty)	7.0

^ahHERG 10 μ M cutoff for active/inactive

4 Diversity Analyses of Natural Products

In addition to the applications of computational methods to study natural products, diversity analysis is one of the most classical and useful applications of cheminformatics. In this section, we describe briefly the sources of natural products with emphasis on the public domain. The reader is referred to a recent chapter of Kirchweger and Rollinger [42] for a more in-depth analysis of this topic. We describe the importance of diversity analysis and discuss representative work on cheminformatic-based analysis of the diversity of natural product collections.

4.1 Overview of Collections of Natural Products

Compound collections are a crucial resource for keeping, searching, mining, and sharing chemical information. Currently, there are several compound databases that enable storing and sharing biological screening data. The relevance of chemical datasets to drug discovery projects has been discussed in detail elsewhere [130]. Interestingly, Clark et al. published initiatives in different countries to promote collaboration in drug discovery projects with research groups in academia [131]. In addition to commercial sources of compounds for computational screening,

there are publicly available large compound databases annotated with biological activity. Representative resources in this regard are ChEMBL, PubChem, and Binding Database, collectively reviewed by Nicola et al. [132]. Of note, as recently commented by Saldívar-González et al. [122], databases annotated with information of the bioactivity profile against one or several biological endpoints are useful for multiple applications including analysis of polypharmacology and structure multiple-activity relationships [133], characterization of activity landscapes [134] and the reexamination of the currently explored chemical space (vide infra).

In 2012, the first databases of natural products available in the public domain at that time were reviewed by Yongye et al. [135]. Six years ago, there were approximately five databases publicly available containing between 560 and 89,000 molecules. Today, many more databases are available with over 250,000 natural products in the public domain as reviewed in the excellent report of Chen et al. [136]. A significant number of natural product resources are built and maintained by academic groups and non-for-profit initiatives. A classic example is the TCM database@Taiwan [137]. Based on this database, iScreen was developed. This is a web server for docking TCM followed by customized de novo drug design [138]. Another example of a previous academic effort is the development of the UNPD [139]. Unfortunately, at the time of writing UNPD is not available. There are other compound collections that are focused on specific geographical regions. A few examples include the NuBBE database that is a collection representative of the Brazilian biodiversity [140, 141]. In turn, the AfroDb collection [142] is an initiative that collects information on the constituents of African medicinal plants, and contains around 1000 three-dimensional structures. The same group developed the ConMedNP collection [143]. Very recently, the VIETHERB database was made available as a compound collection for Vietnamese plant species [144]. In Mexico, Esquivel et al. are building a comprehensive database of natural products that have been published by the Institute of Chemistry of the National Autonomous University of Mexico (UNAM). This database is called UNIQUIM (<http://uniquim.iquimica.unam.mx>). Another initiative from an academic group of the same institution is constructing the BIOFACQUIM database. Currently, BIOFACQUIM contains 423 compounds mostly isolated from Mexican plants and fungi [14]. A comprehensive review of other natural product collections and resources available to the public has been prepared by Chen et al. [136].

4.2 *Design of Nature-Inspired Compound Collections*

In addition to existing collections of natural products, compounds of natural origin have inspired the synthesis of natural product datasets. This comes from the apparent, previously mentioned misapprehension using combinatorial chemistry, as the chemical diversity of the collections made was low [11]. To improve this, natural product scaffolds have been suggested as novel means to access uncharted regions of therapeutic and chemical space [9].

For example, Stratton et al. provided a comprehensive comparison of the chemical space of natural products and drugs [145]. This study highlighted the inherent complexity of natural products as the main tool to effectively optimize lead compounds. A similar observation had previously been suggested in a series of studies by Lovering et al. which tackled the issue of molecular attrition, because of low complexity or “flat molecules” as leads [146, 147]. In addition, the use of natural product scaffolds may provide other advantages, such as the improvement of pharmacokinetic properties, intellectual property [148], and even prodrug design [149].

A noteworthy example of a cheminformatics tool to drive biology-oriented synthesis is Scaffold Hunter [150]. Originally envisioned as a visualization tool, it has overgrown its original purpose allowing further types of analysis. A prominent feature is the so-called Periodic Table of Natural Products, which conducts structural deconvolution to provide vantage points for synthesis routes. Successful cases using this method include 11 β -hydroxysteroid dehydrogenase, 5-lipoxygenase, phosphatase, and kinase inhibitors [151].

4.3 *Concept and Importance of Diversity Analysis*

The continued increase in the number of compounds available in compound databases has led to the notion of chemical space [152] and makes necessary to characterize the content and diversity of the molecules stored in those collections. Indeed, comparison of the content and overall the contents of the molecular databases is important in assortment design and selection [153] as diversity analysis aids in the assessment of the structural novelty of molecules. Systematic analysis of the diversity and chemical space of compound collections, in particular large collections, usually needs cheminformatic approaches [123].

Approaches to assess the diversity of compound databases can be divided into two main groups that largely depend on the molecular representation [14], namely, graphs and descriptor vectors [21, 154]. Graph methods are employed to conduct structural and sub-structural analysis. These approaches are relatively easy to interpret. Representation using descriptor vectors is commonly used in cheminformatics for database processing, similarity searching, clustering, and developing descriptive and predictive models. The choice of descriptors used to analyze compound datasets—with more than five thousand available thus far—gives rise to different types of chemical spaces as pointed out by Varnek and Baskin [154]. The structural diversity of natural product databases using structural fingerprints, molecular scaffolds, and other representation was published in several reports. Analysis of the chemical space of natural product databases has recently been published [14]. In the next section, we will discuss representative studies with emphasis on the diversity analysis that have appeared most recently.

4.4 Representative Diversity Analysis of Natural Products

Table 4 summarizes examples of cheminformatic analysis of natural product collections and other relevant compound collections that are usually used for reference. The table includes the databases analyzed, and the main structural representations employed. Selected studies are further commented below with a focus on the most recent work carried out.

In 2015, Pascolutti et al. published the generation of fragment screening collections that aim to capture the broad range of molecular recognition building blocks included within natural products as included in the “Dictionary of Natural Products” (DNP; Chapman and Hall/CRC Press, Boca Raton, FL, USA). The structural diversity of the fragment versus a reference non-fragment assortment was analyzed using three complementary approaches, namely, atom function analysis (based on pharmacophore fingerprints), atom type analysis (with radial fingerprints), and scaffold analysis. Among the various conclusions made, Pascolutti et al. found that naturally derived fragments could be used as the starting point for building chemical collections with high diversity for medicinal chemistry projects.

Chen et al. [136] reported recently a comprehensive analysis toward the understanding of the population of the chemical space by currently known and accessible natural products and by individual natural product collections. As stated by the authors, among the relevant results of this work was that the easily accessible natural products have a large diversity and cover regions of medicinally relevant chemical

Table 4 Representative studies of chemical diversity of natural products

Datasets	Descriptors/representation	Refs.
TCM, combinatorial libraries, drugs approved for clinical use, and screening collections	Molecular fingerprints, scaffolds, physicochemical properties	[155]
Natural products, human metabolites, bioactive compounds, clinical candidates, and drugs	Topological and physicochemical	[156]
Fragment-sized and no fragment-sized natural products	Pharmacophore and radial fingerprints, and molecular scaffolds	[157]
Eighteen virtual and nine existing natural product libraries. As reference, the “Dictionary of Natural Products” was used	Physicochemical properties	[136]
Cyanobacteria, fungi metabolites, marine, purified natural product screening compounds, TCM, NuBBE _{DB} , UNPD. As reference, semi-synthetic and approved drugs were used	Molecular fingerprints, scaffolds, physicochemical properties; drug-, extended drug-, lead-, fragment-, PPI-like, and PAINS profiling; molecular complexity	[122]
BIOFACQUIM, NuBBE _{DB} , TCM. As reference, approved drugs were used	Molecular fingerprints, scaffolds, physicochemical properties	[14]

space. In some instances, the authors observed a significant difference in the coverage of the chemical space of different classes and individual datasets of natural products.

Saldivar-González et al. reported a comprehensive cheminformatic characterization of seven natural product databases inclusive of cyanobacterial, fungal metabolites, marine, purified natural product screening compounds, TCM, NuBBE, and UNPD databases [122]. As references, a semi-synthetic compound collection and a set of drugs approved for clinical use were employed. The datasets were profiled and compared using a number of different and complementary representations and descriptors, namely, molecular fingerprints of different design (Extended Connectivity fingerprints radius two and Molecular Access System (MACCS) keys), scaffolds, and six physicochemical properties of pharmaceutical interest. In addition, the chemical databases were profiled using empirical rules that have been developed to classify drug-, extended drug-, lead-, fragment-, PPI-like, and PAINS compounds. Finally, the datasets were profiled using two descriptors associated with molecular complexity: fraction of carbon atoms with sp^3 hybridization (FC sp^3) and the fraction of chiral carbons (FCC). Among the conclusions, it was found that the NuBBE database, the main focus of this work, had a restrained chemical space, with the majority within the region of the drug-like physicochemical properties. It was also concluded that the main source of diversity in the compounds in NuBBE database was driven by the side chains. Overall, the results were supportive of a large number of molecules in NuBBE database being promising sources of lead molecules for medicinal chemistry and drug discovery projects [122].

Recently, Pilon-Jiménez et al. discussed the collection and first diversity analysis of BIOFACQUIM, a database of natural products isolated from organisms in Mexico [158]. In that work, the authors characterize the diversity of BIOFACQUIM using molecular fingerprints (MACCS keys), molecular scaffolds, and six drug-like physicochemical properties, namely, molecular weight, topological surface area, number of hydrogen bond donors and acceptors, number of rotatable bonds and the *n*-octanol/water partition coefficient, $\log P$. BIOFACQUIM was compared to other natural product and reference databases such as NuBBE, TCM, and approved drugs. It was found that BIOFACQUIM and AfroDb are diverse in terms of scaffolds, but both have relatively low fingerprint diversity. It was also concluded that AfroDb is more diverse than BIOFACQUIM, in terms of relevant physicochemical properties. In contrast, the set of approved drugs had a medium diversity based on fingerprints and relatively low diversity using the scaffolds. In turn, TCM had the largest scaffold and fingerprint diversity, relative to the datasets compared in that work [14].

4.4.1 Global Analysis of Chemical Diversity

As explained above, chemical representation and descriptors are at the core of diversity analysis and basically any cheminformatic application [114]. Therefore, the perception of the chemical space and assessment of the diversity of a compound

collection in general is relative to the molecular representation. In order to reduce (although not eliminate entirely) the dependence of the diversity with molecular representation, it has been proposed to use a consensus approach through the assessment of the global diversity using Consensus Diversity Plots (CDPs) [159]. Consensus Diversity Plots are two-dimensional graphs to represent simultaneously four diversities (typically fingerprint-based, scaffold, whole molecular properties—associated with drug-like characteristics, and size of the database). Consensus Diversity Plots have been employed to characterize quantitatively the total or global diversity of fungal metabolites [160], natural products from Panama [161], from Brazil as available in NuBBE [122], and from Mexico (as deposited in the BIOFACQUIM database) [14].

Consensus Diversity Plots have also been used to compare the diversity of food chemicals to other datasets [162]. There is a free online server where any user can generate CDPs for their own collections [159]. The server is available through D-TOOLS (www.difacquim.com/d-tools/). To exemplify a CDP, Fig. 5 shows a comparison of the total diversity of the current version of BIOFACQUIM dataset (vide supra) with seven reference datasets [157]. The CDP compares the databases considering as basis of diversity a molecular fingerprint typically used to assess structural diversity (MACCS keys), molecular scaffolds, and the six physicochemical properties SlogP, TPSA, MW, RB, HBD, and HBA (vide supra). The median of the distribution of the MACCS keys (166-bits)/Tanimoto Similarity of each dataset is represented on the x -axis (lower values denote higher fingerprint-based diversity). The y -axis measures the scaffold diversity of each set as the area under the scaffold recovery curve [158]; here lower values denote higher scaffold diversity (where the highest diversity would be an area under the curve of 0.5 [163]). The property-based diversity is represented with the Euclidean distance of the scaled properties, inserted into the maps using a continuous color scale: a darker blue color indicates lower diversity while lighter blue denotes higher property diversity. The relative size of each dataset is mapped with different sizes of the data points, with smaller data points indicating datasets with fewer numbers of molecules. Thus, the CDP indicates, for instance, that BIOFACQUIM and cyanobacteria metabolites have, overall, the lowest scaffold and fingerprint-based diversity (among the datasets compared). Considering the diversity based on physicochemical properties, cyanobacteria metabolites have a larger diversity than compounds in BIOFACQUIM (as indicated by a lighter blue data point). The CDP further indicated that the set of drugs approved for clinical use have a high scaffold and fingerprint-based diversity (as noted for other CDPs, the set of approved drugs tend to have high global diversity [159, 162, 164]).

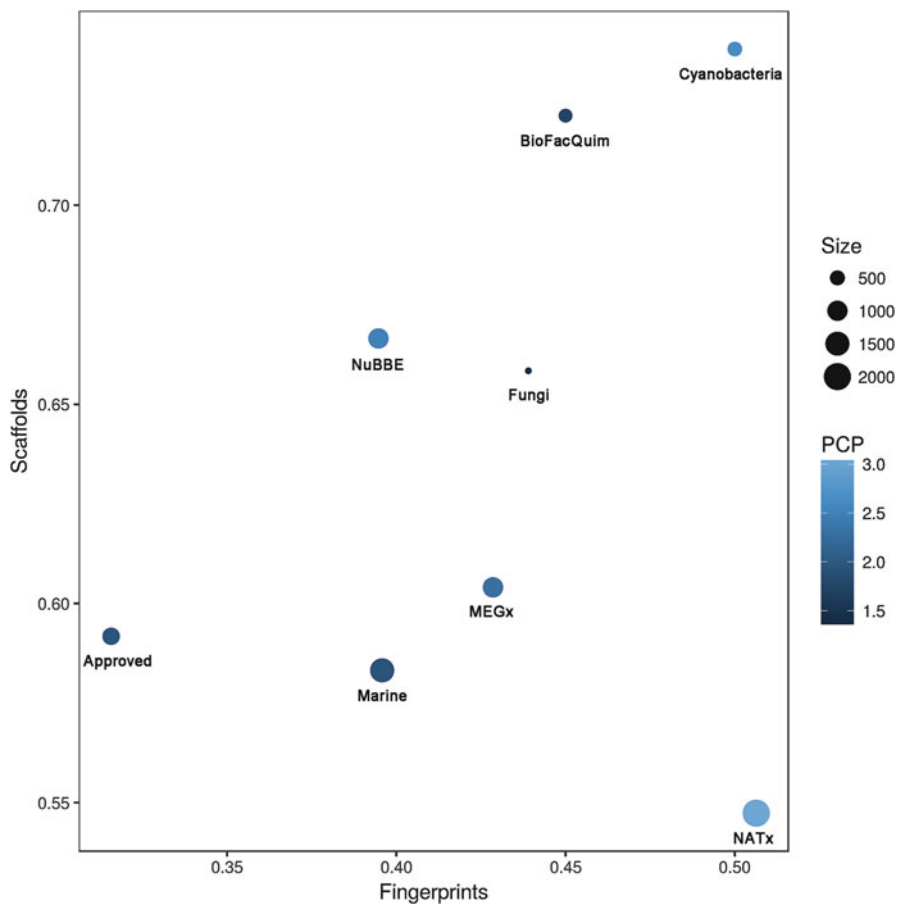


Fig. 5 Consensus Diversity Plot comparing the global diversity of BIOFACQUIM with other natural product databases. The structural diversity (fingerprint diversity) was calculated with the median Tanimoto coefficient of MACCS keys fingerprints is plotted on the *x*-axis. The scaffold diversity of each database was defined as the area under the curve (AUC) of the respective scaffold recovery curves, and it is represented on the *y*-axis. The diversity based on physicochemical properties (PCP) was calculated with the Euclidean distance of six scaled properties (SlogP, TPSA, MW, RB, HBD, and HBA) and is shown in a color scale. The distance is represented with a continuous color scale from light blue (more diverse) to dark blue (less diverse). The relative size of the dataset is represented with the size of the data point: smaller data points indicate compound datasets with fewer molecules

5 Conclusions and Future Directions

Natural products retain a fundamental role in the drug discovery process, despite the implicit difficulties involved. Nonetheless, the industrial setting has favored other approaches leaving such endeavors to academia. With the emergence of

multidisciplinary studies, natural products have seen a renaissance. In this sense, in silico methods provide flexible tools to analyze screens and describe in a qualitative and quantitative basis the diversity, presumptive activity, and even the potential toxicity of natural products.

With several instances of success across different targets, it would seem that natural product research driven by computational methods is “a match made in heaven.” Still, some aspects of computational methodologies cannot be applied “as is,” due to current limitations of the techniques and algorithms. This has had a positive impact in the computational field, stimulating the development of more robust protocols and methods or even a focus toward natural product modeling as a whole. As discussed in this chapter, the availability of new and improved algorithms has led to the development and implementation of a plethora of applications that range from the collection of data to the in silico profiling and screening of natural products. In this sense, the overall projection of computational-based natural product research will continue to thrive, given the increasing number of data sources and the array of metabolites that remain unexplored.

Hence, perspectives on this field regard the construction and optimization of proper databases to enhance fragment-based campaigns and the expansion of chemical space. These include improvement of cheminformatic filters for the identification of activity cliffs.

Acknowledgments Fernando Prieto-Martínez is grateful for a Ph.D. scholarship from the Consejo Nacional de Ciencia y Tecnología (CONACyT) No. 660465/576637. The authors also thank the Programa de Nuevas Alternativas de Tratamiento para Enfermedades Infecciosas (NUATEI-IIB-UNAM). José Medina-Franco acknowledges the School of Chemistry of the Universidad Nacional Autónoma de México (UNAM), the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) grant number IA203718, UNAM and the Consejo Nacional de Ciencia y Tecnología grant number 282785. Fernando Prieto-Martínez and José Medina-Franco also thank Dirección General de Cómputo y de Tecnologías de Información y Comunicación (DGTIC), project grant LANCAD-UNAM-DGTIC-335 for the computational resources to use Miztli supercomputer at UNAM. The authors thank Fernanda I. Saldívar-González for providing the datasets on natural products used to compute the toxicity profile, Dr. Sharon Luna for assisting in the analysis of the toxicity data, and Edgar López-López for helpful discussions.

References

1. Perry NSL, Bollen C, Perry EK, Ballard C (2003) *Salvia* for dementia therapy: review of pharmacological activity and pilot tolerability clinical trial. *Pharmacol Biochem Behav* 75:651
2. Astudillo-Vázquez A, Dávalos Valle H, De Jesús L, Herrera G, Navarrete A (2008) Investigation of *Altermanthera repens* and *Bidens odorata* on gastrointestinal disease. *Fitoterapia* 79:577
3. Baum SS, Hill R, Rommelspacher H (1998) Effect of kava extract and individual kavapyrones on neurotransmitter levels in the nucleus accumbens of rats. *Prog Neuro-Psychopharmacol Biol Psychiatry* 22:1105

4. Chavkin C (2003) Salvinorin A, an active component of the hallucinogenic sage *Salvia divinorum* is a highly efficacious opioid receptor agonist: structural and functional considerations. *J Pharmacol Exp Ther* 308:1197
5. Öztürk Y, Aydın S, Beis R, Başer KH, Berberoğlu H (1996) Effects of *Hypericum perforatum* L. and *Hypericum calycinum* L. extracts on the central nervous system in mice. *Phytomedicine* 3:139
6. Dias DA, Urban S, Roessner U (2012) A historical overview of natural products in drug discovery. *Metabolites* 2:303
7. Beutler JA (2009) Natural products as a foundation for drug discovery. *Curr Protoc Pharmacol* 46:9
8. Harvey AL (2008) Natural products in drug discovery. *Drug Discov Today* 13:894
9. Ortholand JY, Ganesan A (2004) Natural products and combinatorial chemistry: back to the future. *Curr Opin Chem Biol* 8:271
10. Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, Green DV, Hertzberg RP, Janzen WP, Paslay JW, Schopfer U, Sittampalam GS (2011) Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* 10:188
11. Ganesan A (2004) Natural products as a hunting ground for combinatorial chemistry. *Curr Opin Biotechnol* 15:584
12. Cragg GM, Newman DJ (2013) Natural products: a continuing source of novel drug leads. *Biochim Biophys Acta, Gen Subj* 1830:3670
13. Pereira F, Aires-de-Sousa J (2018) Computational methodologies in the exploration of marine natural product leads. *Mar Drugs* 16:236
14. Saldívar-González FI, Pilón-Jiménez BA, Medina-Franco JL (2018) Chemical space of naturally occurring compounds. *Phys Sci Rev.* <https://doi.org/10.1515/psr-2018-0103>
15. Thomford NE, Senteheane DA, Rowe A, Munro D, Seele P, Maroyi A, Dzobo K (2018) Natural products for drug discovery in the 21st century: innovations for novel drug discovery. *Int J Mol Sci* 19:1578
16. González-Medina M, Naveja JJ, Sánchez-Cruz N, Medina-Franco JL (2017) Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. *RSC Adv* 7:54153
17. Baell JB, Nissink JWM (2018) Seven year itch: pan-assay interference compounds (PAINS) in 2017 – utility and limitations. *ACS Chem Biol* 13:36
18. Willett P (2011) Chemoinformatics: a history. *Wiley Interdiscip Rev Comput Mol Sci* 1:46
19. Engel T (2006) Basic overview of chemoinformatics. *J Chem Inf Model* 46:2267
20. Opassi G, Gesù A, Massarotti A (2018) The hitchhiker's guide to the chemical-biological galaxy. *Drug Discov Today* 23:565
21. Maggiora GM, Shanmugasundaram V (2011) Molecular similarity measures. *Humana, Totowa, NJ*, p 39
22. Lill MA (2007) Multi-dimensional QSAR in drug discovery. *Drug Discov Today* 12:1013
23. Prieto-Martínez FD, Medina-Franco JL (2018) Molecular docking: current advances and challenges. *TIP Rev Espec Ciencias Químico-Biológicas* 25:65
24. Schlick T (2010) Molecular dynamics: basics. In: *Molecular modeling and simulation. An interdisciplinary guide*, 2nd edn. Springer, New York, p 425
25. Parenti MD, Rastelli G (2012) Advances and applications of binding affinity prediction methods in drug discovery. *Biotechnol Adv* 30:244
26. Lavecchia A, Giovanni C (2013) Virtual screening strategies in drug discovery: a critical review. *Curr Med Chem* 20:2839
27. Rollinger JM, Stuppner H, Langer T (2008) Virtual screening for the discovery of bioactive natural products. *Prog Drug Res* 65:211
28. Ma D-L, Chan DS-H, Leung C-H (2011) Molecular docking for virtual screening of natural product databases. *Chem Sci* 2:1656
29. Kubinyi H (2008) QSAR: Hansch analysis and related approaches. VCH, Weinheim

30. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2012) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 64:4
31. Leeson PD (2015) Molecular inflation, attrition and the rule of five. *Adv Drug Deliv Rev* 101:22
32. Deshpande M, Kuramochi M, Karypis G (2007) Data mining algorithms for virtual screening of bioactive compounds. *Springer Optim Appl* 7:59
33. Rathi PC, Ludlow RF, Hall RJ, Murray CW, Mortenson PN, Verdonk ML (2017) Predicting “hot” and “warm” spots for fragment binding. *J Med Chem* 60:4036
34. Cerqueira NMFS, Gesto D, Oliveira EF, Santos-Martins D, Brás NF, Sousa SF, Fernandes PA, Ramos MJ (2015) Receptor-based virtual screening protocol for drug discovery. *Arch Biochem Biophys* 582:56
35. Wingert BM, Camacho CJ (2018) Improving small molecule virtual screening strategies for the next generation of therapeutics. *Curr Opin Chem Biol* 44:87
36. McInnes C (2007) Virtual screening strategies in drug discovery. *Curr Opin Chem Biol* 11:494
37. Spyrikis F, Cavasotto CN (2015) Open challenges in structure-based virtual screening: receptor modeling, target flexibility consideration and active site water molecules description. *Arch Biochem Biophys* 583:105
38. Bajusz D, Rácz A, Héberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* 7:20
39. Tanrikulu Y, Krüger B, Proschak E (2013) The holistic integration of virtual screening in drug discovery. *Drug Discov Today* 18:358
40. Cruz-Monteagudo M, Medina-Franco JL, Pérez-Castillo Y, Nicolotti O, Cordeiro MN, Borges F (2014) Activity cliffs in drug discovery: Dr. Jekyll or Mr. Hyde? *Drug Discov Today* 19:1069
41. Kukul A (2011) Consensus virtual screening approaches to predict protein ligands. *Eur J Med Chem* 46:4661
42. Kirchweger B, Rollinger JM (2018) Virtual screening for the discovery of active principles from natural products. In: *Natural products as source of molecules with therapeutic potential*, p 333
43. Scotti L, Bezerra Mendonca FJ, Ribeiro FF, Tavares JF, da Silva MS, Barbosa Filho JM, Scotti MT (2018) Natural product inhibitors of topoisomerases: review and docking study. *Curr Protein Pept Sci* 19:275
44. Jenkins JL, Bender A, Davies JW (2006) In silico target fishing: predicting biological targets from chemical structure. *Drug Discov Today Technol* 3:413
45. Szyf M (2015) Epigenetics, a key for unlocking complex CNS disorders? Therapeutic implications. *Eur Neuropsychopharmacol* 25:682
46. Biswas S, Rao CM (2018) Epigenetic tools (the writers, the readers and the erasers) and their implications in cancer therapy. *Eur J Pharmacol* 837:8
47. Schwenk RW, Vogel H, Schürmann A (2013) Genetic and epigenetic control of metabolic health. *Mol Metab* 2:337
48. Paneni F, Costantino S, Volpe M, Lüscher TF, Cosentino F (2013) Epigenetic signatures and vascular risk in type 2 diabetes: a clinical perspective. *Atherosclerosis* 230:191
49. Wilting RH, Dannenberg J-H (2012) Epigenetic mechanisms in tumorigenesis, tumor cell heterogeneity and drug resistance. *Drug Resist Updat* 15:21
50. Miousse IR, Currie R, Datta K, Ellinger-Ziegelbauer H, French JE, Harrill AH, Koturbash I, Lawton M, Mann D, Meehan RR, Moggs JG, O'Lone R, Rasoulpour RJ, Pera RA, Thompson K (2015) Importance of investigating epigenetic alterations for industry and regulators: an appraisal of current efforts by the Health and Environmental Sciences Institute. *Toxicology* 335:11
51. Wegner M, Neddermann D, Piorunski-Stolzmann M, Jagodzinski PP (2014) Role of epigenetic mechanisms in the development of chronic complications of diabetes. *Diabetes Res Clin Pract* 105:164

52. Cabaye A, Nguyen KT, Liu L, Pande V, Schapira M (2015) Structural diversity of the epigenetics pocketome. *Proteins Struct Funct Bioinf* 83:1316
53. Pande V (2016) Understanding the complexity of epigenetic target space. *J Med Chem* 59:1299
54. Priestley CC, Anderton M, Doherty AT, Duffy P, Mellor HR, Powella H, Roberts R (2012) Epigenetics – relevance to drug safety science. *Toxicol Res* 1:23
55. Shortt J, Ott CJ, Johnstone RW, Bradner JE (2017) A chemical probe toolbox for dissecting the cancer epigenome. *Nat Rev Cancer* 17:160
56. Fischle W, Schwarzer D (2016) Probing chromatin-modifying enzymes with chemical tools. *ACS Chem Biol* 11:689
57. Singh M, Kaur M, Silakari O (2014) Flavones: an important scaffold for medicinal chemistry. *Eur J Med Chem* 84:206
58. Vasantha Rupasinghe HP, Nair SVG, Robinson RA (2014) Chemopreventive properties of fruit phenolic compounds and their possible mode of actions, 1st edn. Elsevier, Amsterdam
59. Ferguson FM, Fedorov O, Chaikwad A, Philpott M, Muniz JR, Felletar I, von Delft F, Heightman T, Knapp S, Abell C, Ciulli A (2013) Targeting low-druggability bromodomains: fragment based screening and inhibitor design against the BAZ2B bromodomain. *J Med Chem* 56:10183
60. Prinjha RK, Witherington J, Lee K (2012) Place your BETs: the therapeutic potential of bromodomains. *Trends Pharmacol Sci* 33:146
61. Prieto-Martínez FD, Fernández-de Gortari E, Méndez-Lucio O, Medina-Franco JL (2016) A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. *RSC Adv* 6:56225
62. Zhao H, Gartenmann L, Dong J, Spiliotopoulos D, Caffisch A (2014) Discovery of BRD4 bromodomain inhibitors by fragment-based high-throughput docking. *Bioorg Med Chem Lett* 24:2493
63. Hoffer L, Voitovich YV, Raux B, Carrasco K, Muller C, Fedorov AY, Derviaux C, Amouric A, Betzi S, Horvath D, Varnek A, Collette Y, Combes S, Roche P, Morelli X (2018) Integrated strategy for lead optimization based on fragment growing: the diversity-oriented-target-focused-synthesis approach. *J Med Chem* 61:5719
64. Tanaka M, Roberts JM, Seo H-S, Souza A, Paulk J, Scott TG, DeAngelo SL, Dhe-Paganon S, Bradner JE (2016) Design and characterization of bivalent BET inhibitors. *Nat Chem Biol* 12:1089
65. Spiliotopoulos D, Caffisch A (2014) Molecular dynamics simulations of bromodomains reveal binding-site flexibility and multiple binding modes of the natural ligand acetyl-lysine. *Isr J Chem* 54:1084
66. Prieto-Martínez FD, Medina-Franco JL (2018) Charting the bromodomain BRD4: towards the identification of novel inhibitors with molecular similarity and receptor mapping. *Lett Drug Des Discov* 15:1
67. Tarallo V, Lepore L, Marcellini M, Dal Piaz F, Tudisco L, Ponticelli S, Lund FW, Roepstorff P, Orlandi A, Pisano C, De Tommasi N, De Falco S (2011) The biflavonoid amentoflavone inhibits neovascularization preventing the activity of proangiogenic vascular endothelial growth factors. *J Biol Chem* 286:19641
68. Liu H, Yue Q, He S (2017) Amentoflavone suppresses tumor growth in ovarian cancer by modulating Skp2. *Life Sci* 189:96
69. Dhananjayan K (2015) Molecular docking study characterization of rare flavonoids at the Nac-binding site of the first bromodomain of BRD4 (BRD4 BD1). *J Cancer Res* 2015:1
70. Raj U, Kumar H, Varadwaj PK (2016) Molecular docking and dynamics simulation study of flavonoids as BET bromodomain inhibitors. *J Biomol Struct Dyn* 1102:1
71. Waterman MJ, Nugraha AS, Hendra R, Ball GE, Robinson SA, Keller PA (2017) Antarctic moss biflavonoids show high antioxidant and ultraviolet-screening activity. *J Nat Prod* 80:2224

72. Bharatham N, Slavish PJ, Young BM, Shelat AA (2018) The role of ZA channel water-mediated interactions in the design of bromodomain-selective BET inhibitors. *J Mol Graph Model* 81:197
73. Jung M, Philpott M, Müller S, Schulze J, Badock V, Eberspächer U, Moosmayer D, Bader B, Schmees N, Fernández-Montalván A, Haendler B (2014) Affinity map of bromodomain protein 4 (BRD4) interactions with the histone H4 tail and the small molecule inhibitor JQ1. *J Biol Chem* 289:9304
74. Kharenko OA, Gesner EM, Patel RG, Norek K, White A, Fontano E, Suto RK, Young PR, McLure KG, Hansen HC (2016) RVX-297 — a novel BD2 selective inhibitor of BET bromodomains. *Biochem Biophys Res Commun* 477:62
75. Prieto-Martínez FD, Medina-Franco JL (2018) Flavonoids as putative epi-modulators: insight into their binding mode with BRD4 bromodomains using molecular docking and dynamics. *Biomolecules* 8:61
76. Shadrack WR, Slavish PJ, Chai SC, Waddell B, Connelly M, Low JA, Tallant C, Young BM, Bharatham N, Knapp S, Boyd VA, Morfouace M, Roussel MF, Chen T, Lee RE, Kiplin Guy R, Shelat AA, Potter PM (2018) Exploiting a water network to achieve enthalpy-driven, bromodomain-selective BET inhibitors. *Bioorg Med Chem* 26:25
77. Guha M (2015) HDAC inhibitors still need a home run, despite recent approval. *Nat Rev Drug Discov* 14:225
78. Robert C, Rassool FV (2012) HDAC inhibitors. In: *Histone deacetylase inhibitors as cancer therapeutics*, 1st edn. Elsevier, Amsterdam, p 87
79. Zhu S, Dong Z, Ke X, Hou J, Zhao E, Zhang K, Wang F, Yang L, Xiang Z, Cui H (2018) The roles of sirtuins family in cell metabolism during tumor development. *Semin Cancer Biol*. <https://doi.org/10.1016/j.semcancer.2018.11.003>
80. Jing H, Lin H (2015) Sirtuins in epigenetic regulation. *Chem Rev* 115:2350
81. Wątroba M, Dudek I, Skoda M, Stangret A, Rzodkiewicz Z, Szukiewicz D (2017) Sirtuins, epigenetics and longevity. *Ageing Res Rev* 40:11
82. Dai H, Sinclair DA, Ellis JL, Steegborn C (2018) Sirtuin activators and inhibitors: promises, achievements, and challenges. *Pharmacol Ther* 188:140
83. Ueda H, Nakajima H, Hori Y, Fujita T, Nishimura M, Goto T, Okuhara M (1994) FR901228, a novel antitumor bicyclic depsipeptide produced by *Chromobacterium violaceum* No. 968. II. Structure determination. *J Antibiot* 47:301
84. Robey RW, Chakraborty AR, Basseville A, Luchenko V, Bahr J, Zhan Z, Bates SE (2011) Histone deacetylase inhibitors: emerging mechanisms of resistance. *Mol Pharmaceutics* 8:2021
85. Konstantinopoulos PA, Vondoros GP, Papavassiliou AG (2006) FK228 (depsipeptide): a HDAC inhibitor with pleiotropic antitumor activities. *Cancer Chemother Pharmacol* 58:711
86. VanderMolen KM, McCulloch W, Pearce CJ, Oberlies NH (2011) Romidepsin (Istodax, NSC 630176, FR901228, FK228, depsipeptide): a natural product recently approved for cutaneous T-cell lymphoma. *J Antibiot* 64:525
87. Cherblanc FL, Davidson RWM, Di Fruscia P, Srimongkolpithak N, Fuchter MJ (2013) Perspectives on natural product epigenetic modulators in chemical biology and medicine. *Nat Prod Rep* 30:605
88. Neugebauer RC, Uchieczowska U, Meier R, Hruby H, Valkov V, Verdin E, Sippl W, Jung M (2008) Structure-activity studies on splitomicin derivatives as sirtuin inhibitors and computational prediction of binding mode. *J Med Chem* 51:1203
89. Kokkonen P, Mellini P, Nyrhilä O, Rahnasto-Rilla M, Suuronen T, Kiviranta P, Huhtiniemi T, Poso A, Jarho E, Lahtela-Kakkonen M (2014) Quantitative insights for the design of substrate-based SIRT1 inhibitors. *Eur J Pharm Sci* 59:12
90. Sun Y, Zhou H, Zhu H, Leung SW (2016) Ligand-based virtual screening and inductive learning for identification of SIRT1 inhibitors in natural products. *Sci Rep* 6:1

91. Wang Y, Liang X, Chen Y, Zhao X (2016) Screening SIRT1 activators from medicinal plants as bioactive compounds against oxidative damage in mitochondrial function. *Oxidative Med Cell Longev* 2016:1
92. Karaman B, Alhalabi Z, Swyter S, Mihigo SO, Andrae-Marobela K, Jung M, Sippl W, Ntie-Kang F (2018) Identification of bichalcones as sirtuin inhibitors by virtual screening and in vitro testing. *Molecules* 23:1
93. Wang Y, He J, Liao M, Hu M, Li W, Ouyang H, Wang X, Ye T, Zhang Y, Ouyang L (2019) An overview of sirtuins as potential therapeutic target: structure, function and modulators. *Eur J Med Chem* 161:48
94. Rahnasto-Rilla M, Tyni J, Huovinen M, Jarho E, Kulikowicz T, Ravichandran S, A Bohr V, Ferrucci L, Lahtela-Kakkonen M, Moaddel R (2018) Natural polyphenols as sirtuin 6 modulators. *Sci Rep* 8:1
95. Religa AA, Waters AP (2012) Sirtuins of parasitic protozoa: in search of function(s). *Mol Biochem Parasitol* 185:71
96. Mittal N, Muthuswami R, Madhubala R (2017) The mitochondrial SIR2 related protein 2 (SIR2RP2) impacts *Leishmania donovani* growth and infectivity. *PLoS Negl Trop Dis* 1: e0005590
97. Ritagliati C, Alonso VL, Manarin R, Cribb P, Serra EC (2015) Overexpression of cytoplasmic TcSIR2RP1 and mitochondrial TcSIR2RP3 impacts on *Trypanosoma cruzi* growth and cell invasion. *PLoS Negl Trop Dis* 9:1
98. Kadam RU, Tavares J, Kiran VM, Cordeiro A, Ouassii A, Roy N (2008) Structure function analysis of *Leishmania* sirtuin: an ensemble of in silico and biochemical studies. *Chem Biol Drug Des* 71:501
99. Soares MBP, Silva CV, Bastos TM, Guimaraes ET, Figueira CP, Smirlis D, Azevedo WF Jr (2012) Anti-*Trypanosoma cruzi* activity of nicotinamide. *Acta Trop* 122:224
100. Rose NR, Klose RJ (2014) Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta — Gene Regul Mech* 1839:1362
101. Liu Y, Liu K, Qin S, Xu C, Min J (2014) Epigenetic targets and drug discovery: Part 1: histone methylation. *Pharmacol Ther* 143:275
102. Zhang J, Zheng YG (2016) SAM/SAH analogs as versatile tools for SAM-dependent methyltransferases. *ACS Chem Biol* 11:583
103. Zheng W, Ibáñez G, Wu H, Blum G, Zeng H, Dong A, Li F, Hajian T, Allali-Hassani A, Amaya MF, Siarheyeva A, Yu W, Brown PJ, Schapira M, Vedadi M, Min J, Luo M (2012) Sinefungin derivatives as inhibitors and structure probes of protein lysine methyltransferase SETD2. *J Am Chem Soc* 134:18004
104. Fernández-de Gortari E, Medina-Franco JL (2015) Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases. *RSC Adv* 5:87465
105. Marzag H, Warnault P, Bougrin K, Martinet N, Benhida R (2014) Natural polyphenols as potent inhibitors of DNA methyltransferases, 1st edn. Elsevier, Amsterdam
106. Maugeri A, Barchitta M, Mazzone MG, Giuliano F, Basile G, Agodi A (2018) Resveratrol modulates SIRT1 and DNMT functions and restores LINE-1 methylation levels in ARPE-19 cells under oxidative stress and inflammation. *Int J Mol Sci* 19:1
107. Aldawsari FS, Aguayo-Ortiz R, Kapilashrami K, Yoo J, Luo M, Medina-Franco JL, Velázquez-Martínez CA (2016) Resveratrol-salicylate derivatives as selective DNMT3 inhibitors and anticancer agents. *J Enzyme Inhib Med Chem* 31:695
108. Weng JR, Lai IL, Yang HC, Lin CN, Bai LY (2014) Identification of kazinol Q, a natural product from Formosan plants, as an inhibitor of DNA methyltransferase. *Phytother Res* 28:49
109. Parasuraman S (2011) Toxicological screening. *J Pharmacol Pharmacother* 2:74
110. Gleeson MP, Modi S, Bender A, Robinson RL, Kirchmair J, Promkatkaew M, Hannongbua S, Glen RC (2012) The challenges involved in modeling toxicity data in silico: a review. *Curr Pharm Des* 18:1266

111. Sosnin S, Karlov D, Tetko IV, Fedorov MV (2018) A comparative study of multitask toxicity modeling on a broad chemical space. *J Chem Inf Model*. <https://doi.org/10.1021/acs.jcim.8b00685>
112. Hamadache M, Amrane A, Benkortbi O, Hanini S, Khaouane L, Moussa CS (2017) Environmental toxicity of pesticides, and its modeling by QSAR approaches, vol 471. Springer, Cham, Switzerland
113. Peters JU (2013) Polypharmacology – foe or friend? *J Med Chem* 56:8955
114. Maggiora G, Gokhale V (2017) A simple mathematical approach to the analysis of polypharmacology and polyspecificity data. *F1000Research* 6:788
115. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* 53:2719
116. Stork C, Wagner J, Friedrich N-O, de Bruyn KC, Šicho M, Kirchmair J (2018) Hit Dexter: a machine-learning model for the prediction of frequent hitters. *ChemMedChem* 13:564
117. Baell JB (2016) Feeling Nature's PAINS: natural products, natural product drugs, and Pan Assay Interference Compounds (PAINS). *J Nat Prod* 79:616
118. Arvidson KB, Valerio LG, Diaz M, Chanderbhan RF (2008) In silico toxicological screening of natural products. *Toxicol Mech Methods* 18:229
119. Onguéné PA, Simoben CV, Fotso GW, Andrae-Marobela K, Khalid SA, Ngadjui BT, Mbaze LM, Ntie-Kang F (2018) In silico toxicity profiling of natural product compound libraries from African flora with anti-malarial and anti-HIV properties. *Comput Biol Chem* 72:136
120. Ruiz-Rodríguez MA, Vedani A, Flores-Mireles AL, Cháirez-Ramírez MH, Gallegos-Infante JA, González-Laredo RF (2017) In silico prediction of the toxic potential of lupeol. *Chem Res Toxicol* 30:1562
121. Martínez-Mayorga K, Marmolejo-Valencia AF, Cortes-Guzman F, García-Ramos JC, Sánchez-Flores EI, Barroso-Flores J, Medina-Franco JL, Esquivel-Rodríguez B (2017) Toxicity assessment of structurally relevant natural products from Mexican plants with antinociceptive activity toxicity. *J Mex Chem Soc* 61:186
122. Saldívar-González FI, Valli M, Andricopulo AD, da Silva BV, Medina-Franco JL (2019) Chemical space and diversity of the NuBBE database: a chemoinformatic characterization. *J Chem Inf Model* 59:74
123. Medina-Franco JL (2013) Chemoinformatic characterization of the chemical space and molecular diversity of compound libraries. In: *Diversity-oriented synthesis*. Wiley, Hoboken, NJ, p 325
124. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* 52:1757
125. Saqib U, Kelley TT, Panguluri SK, Liu D, Savai R, Baig MS, Schürer SC (2018) Polypharmacology or promiscuity? Structural interactions of resveratrol with its bandwagon of targets. *Front Pharmacol* 9:1201
126. Bisson J, McAlpine JB, Friesen JB, Chen SN, Graham J, Pauli GF (2016) Can invalid bioactives undermine natural product-based drug discovery? *J Med Chem* 59:1671
127. Nelson KM, Dahlin JL, Bisson J, Graham J, Pauli GF, Walters MA (2017) The essential medicinal chemistry of curcumin. *J Med Chem* 60:1620
128. Gavaghan CL, Arnby CH, Blomberg N, Strandlund G, Boyer S (2007) Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J Comput Aided Mol Des* 21:189
129. Kier LD (1985) Use of the Ames test in toxicology. *Regul Toxicol Pharmacol* 5:59
130. Moura Barbosa AJ, Del Rio A (2012) Freely accessible databases of commercial compounds for high-throughput virtual screenings. *Curr Top Med Chem* 12:866
131. Clark RL, Johnston BF, Mackay SP, Breslin CJ, Robertson MN, Harvey AL (2010) The Drug Discovery Portal: a resource to enhance drug discovery from academia. *Drug Discov Today* 15:679

132. Nicola G, Liu T, Gilson MK (2012) Public domain databases for medicinal chemistry. *J Med Chem* 55:6987
133. Saldívar-González FI, Naveja JJ, Palomino-Hernández O, Medina-Franco JL (2017) Getting SMART in drug discovery: cheminformatics approaches for mining structure-multiple activity relationships. *RSC Adv* 7:632
134. Medina-Franco JL, Navarrete-Vázquez G, Méndez-Lucio O (2015) Activity and property landscape modeling is at the interface of cheminformatics and medicinal chemistry. *Future Med Chem* 7:1197
135. Yongye AB, Medina-Franco JL (2012) Data mining of protein-binding profiling data identifies structural modifications that distinguish selective and promiscuous compounds. *J Chem Inf Model* 52:2454
136. Chen Y, Garcia De Lomana M, Friedrich NO, Kirchmair J (2018) Characterization of the chemical space of known and readily obtainable natural products. *J Chem Inf Model* 58:1518
137. Chen CY-C (2011) TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One* 6:e15939
138. Tsai T-Y, Chang K-W, Chen CY-C (2011) iScreen: world's first cloud-computing web server for virtual screening and de novo drug design based on TCM database@Taiwan. *J Comput Aided Mol Des* 25:525
139. Gu J, Gui Y, Chen L, Yuan G, Lu HZ, Xu X (2013) Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One* 8:e62839
140. Valli M, dos Santos RN, Figueira LD, Nakajima CH, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2013) Development of a natural products database from the biodiversity of Brazil. *J Nat Prod* 76:439
141. Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2017) NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci Rep* 7:7215
142. Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, Mbah JA, Mbaze LM, Sippl W, Efang SM (2013) AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PLoS One* 8:e78085
143. Ntie-Kang F, Onguéné PA, Scharfe M, Owono LCO, Megnassan E, Mbaze LM, Sippl W, Efang SM (2014) ConMedNP: a natural product library from central African medicinal plants for drug discovery. *RSC Adv* 4:409
144. Nguyen-Vo T-H, Le T, Pham D, Nguyen TD, Le PH, Nguyen ADT, Nguyen TD, Nguyen TN, Nguyen VA, Do HT, Trinh K, Duong HT, Le LT (2019) VIETHERB: a database for Vietnamese herbal species. *J Chem Inf Model* 59:1
145. Stratton CF, Newman DJ, Tan DS (2015) Cheminformatic comparison of approved drugs from natural product versus synthetic origins. *Bioorg Med Chem Lett* 25:4802
146. Lovering F, Bikker J, Humblet C (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. *J Med Chem* 52:6752
147. Lovering F (2013) Escape from flatland 2: complexity and promiscuity. *Med Chem Commun* 4:515
148. Chen J, Li W, Yao H, Xu J (2015) Insights into drug discovery from natural products through structural modification. *Fitoterapia* 103:231
149. Kumar SV, Saravanan D, Kumar B, Jayakumar A (2014) An update on prodrugs from natural products. *Asian Pac J Trop Med* 7:S54
150. Schäfer T, Kriege N, Humbeck L, Klein K, Koch O, Mutzel P (2017) Scaffold Hunter: a comprehensive visual analytics framework for drug discovery. *J Cheminform* 9:28
151. Rodrigues T (2017) Harnessing the potential of natural products in drug discovery from a cheminformatics vantage point. *Org Biomol Chem* 15:9275
152. Medina-Franco J, Martinez-Mayorga K, Giulianotti M, Houghten RA, Pinilla C (2008) Visualization of the chemical space in drug discovery. *Curr Comput-Aided-Drug Des* 4:322
153. Fitzgerald SH, Sabat M, Geysen HM (2006) Diversity Space and its application to library selection and design. *J Chem Inf Model* 46:1588

154. Varnek A, Baskin II (2011) Chemoinformatics as a theoretical chemistry discipline. *Mol Inform* 30:20
155. López-Vallejo F, Giulianotti MA, Houghten RA, Medina-Franco JL (2012) Expanding the medicinally relevant chemical space with compound libraries. *Drug Discov Today* 17:718
156. Chen H, Engkvist O, Blomberg N, Li J (2012) A comparative analysis of the molecular topologies for drugs, clinical candidates, natural products, human metabolites and general bioactive compounds. *MedChemCommun* 3:312
157. Pascolutti M, Campitelli M, Nguyen B, Pham N, Gorse AD, Quinn RJ (2015) Capturing Nature's diversity. *PLoS One* 10:e0120942
158. Pílon-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL (2019) BIOFACQUIM: a Mexican compound database of natural products. *Biomolecules* 9(1):31
159. González-Medina M, Prieto-Martínez FD, Owen JR, Medina-Franco JL (2016) Consensus diversity plots: a global diversity analysis of chemical libraries. *J Cheminform* 8:63
160. González-Medina M, Owen JR, El-Elimat T, Pearce CJ, Oberlies NH, Figueroa M, Medina-Franco JL (2017) Scaffold diversity of fungal metabolites. *Front Pharmacol* 8:180
161. Olmedo DA, González-Medina M, Gupta MP, Medina-Franco JL (2017) Cheminformatic characterization of natural products from Panama. *Mol Divers* 21:779
162. Naveja JJ, Rico-Hidalgo MP, Medina-Franco JL (2018) Analysis of a large food chemical database: chemical space, diversity, and complexity. *F1000Research* 7:993
163. Medina-Franco JL, Martínez-Mayorga K, Bender A, Scior T (2009) Scaffold diversity analysis of compound datasets using an entropy-based measure. *QSAR Comb Sci* 28:1551
164. González-Medina M, Prieto-Martínez FD, Naveja JJ, Méndez-Lucio O, El-Elimat T, Pearce CJ, Oberlies NH, Figueroa M, Medina-Franco JL (2016) Chemoinformatic expedition of the chemical space of fungal products. *Future Med Chem* 06:1113



Fernando D. Prieto-Martínez obtained his bachelor's degree as an Industrial Pharmaceutical Chemist in 2014 from the National School of Biology of the National Polytechnic Institute (ENCB, IPN; Mexico). His thesis focused on the experimental validation of traditional medicine using plant extracts. That same year he went to the School of Chemistry at National Autonomous University of Mexico (UNAM, Mexico City), where he obtained his master's degree in chemistry. There he had a medicinal chemistry project, focused on the cheminformatics characterization of epigenetic modulators. Currently, he is continuing at the School of Chemistry to obtain his Ph.D. degree, where he is being supervised by Dr. José L. Medina-Franco, and is working on BET bromodomains to develop novel inhibitors of these targets with the use of *in silico* methods. His research interests include molecular dynamics, cheminformatics, and natural products chemistry. He has coauthored nine publications in peer-reviewed journals and four book chapters.



Ulf Norinder received his M.S. in chemical engineering and a Ph.D. degree in chemistry from Chalmers University of Technology (CTH), Gothenburg, Sweden. From 1985 to 2014, he worked as a computational chemist, senior principal scientist, and research fellow in the pharmaceutical industry (Karo Bio AB, AstraZeneca AB and H. Lundbeck A/S), and from 2015 to 2018 as a senior research specialist at Swetox (Karolinska Institutet). He is currently an affiliated researcher at the Department of Computer and Systems Sciences, Stockholm University. His areas of expertise include computer-assisted drug design and pattern recognition with a special emphasis on multivariate data analysis and machine learning.



José L. Medina-Franco received his Ph.D. degree from the National Autonomous University of Mexico (UNAM, Mexico). He was a postdoctoral fellow at the University of Arizona and then joined the Torrey Pines Institute for Molecular Studies in Florida in 2007. In 2013, he moved to the Mayo Clinic and later joined UNAM as full-time Research Professor. He currently leads the DIFACQUIM research group. In 2017, he was named Fellow of the Royal Society of Chemistry (UK). His research interests include the development and application of cheminformatics and molecular modeling methods for bioactive compounds, with an emphasis on drug discovery.

Resources for Chemical, Biological, and Structural Data on Natural Products



Ya Chen, Christina de Bruyn Kops, and Johannes Kirchmair

Contents

1	Introduction	39
2	Virtual Natural Product Databases	40
2.1	Encyclopedic and General Natural Product Databases	45
2.1.1	Dictionary of Natural Products (DNP)	45
2.1.2	AntiBase	45
2.1.3	Reaxys	45
2.1.4	Super Natural II	45
2.1.5	Universal Natural Products Database (UNPD)	46
2.1.6	Natural Product Activity and Species Source (NPASS)	46
2.1.7	Collective Molecular Activities of Useful Plants (CMAUP)	47
2.1.8	Natural Product Atlas	47
2.1.9	Pye et al. Dataset	47
2.1.10	Natural Products Included in the PubChem Substance Database	47
2.1.11	UEFS Natural Products	48
2.2	Databases Focused on Traditional Medicines	48
2.2.1	Traditional Chinese Medicine Database@Taiwan	48
2.2.2	Traditional Chinese Medicine Integrated Database (TCMID 2.0)	48
2.2.3	Yet Another Traditional Chinese Medicine Database (YaTCM)	49
2.2.4	Chemical Database of Traditional Chinese Medicine (Chem-TCM)	49
2.2.5	Herbal Ingredients In Vivo Metabolism Database (HIM)	50
2.2.6	Herbal Ingredients' Targets Database (HIT)	50

Y. Chen · C. de Bruyn Kops

Faculty of Mathematics, Informatics, and Natural Sciences, Department of Computer Science,
Center for Bioinformatics, Universität Hamburg, Hamburg, Germany
e-mail: chen@zbh.uni-hamburg.de; kops@zbh.uni-hamburg.de

J. Kirchmair (✉)

Department of Chemistry, University of Bergen, Bergen, Norway

Computational Biology Unit (CBU), University of Bergen, Bergen, Norway

Faculty of Mathematics, Informatics, and Natural Sciences, Department of Computer Science,
Center for Bioinformatics, Universität Hamburg, Hamburg, Germany
e-mail: johannes.kirchmair@uib.no

© Springer Nature Switzerland AG 2019

A. D. Kinghorn, H. Falk, S. Gibbons, J. Kobayashi, Y. Asakawa, J.-K. Liu (eds.),
Progress in the Chemistry of Organic Natural Products, Vol. 110,
https://doi.org/10.1007/978-3-030-14632-0_2

37

2.2.7	Indian Medicinal Plants, Phytochemistry, and Therapeutics Database (IMPPAT)	50
2.3	Databases Focused on a Specific Habitat or Geographic Region	50
2.3.1	Dictionary of Marine Natural Products (DMNP)	50
2.3.2	MarinLit Database	51
2.3.3	Taiwan Indigenous Plant Database (TIPdb)	51
2.3.4	Northern African Natural Products Database (NANPDB)	51
2.3.5	AfroDb Database	51
2.3.6	South African Natural Compound Database (SANCCDB)	52
2.3.7	African Anticancer Natural Products Library (AfroCancer)	52
2.3.8	African Antimalarial Natural Products Library (AfroMalariaDB)	52
2.3.9	Nuclei of Bioassays, Biosynthesis, and Ecophysiology of Natural Products Database (NuBBEDB)	52
2.3.10	BIOFACQUIM Database	53
2.4	Databases Focused on Specific Organisms	53
2.4.1	<i>Pseudomonas aeruginosa</i> Metabolome Database (PAMDB)	53
2.4.2	StreptomeDB 2.0	53
2.5	Databases Focused on Specific Biological Activities	54
2.5.1	Database of Natural Products for Cancer Gene Regulation (NPCARE)	54
2.5.2	Naturally Occurring Plant-Based Anti-cancer Compound-Activity-Target Database (NPACT)	54
2.5.3	InflamNat Database	54
2.6	Databases Focused on Specific Natural Product Classes	55
2.6.1	Carotenoids Database	55
3	Physical Natural Product Collections	55
3.1	Pure Natural Product Collections	58
3.1.1	Ambinter and Greenpharma	58
3.1.2	AnalytiCon Discovery	59
3.1.3	Chengdu Biopurify Phytochemicals	59
3.1.4	Selleck Chemicals	59
3.1.5	TargetMol Collection	59
3.1.6	MedChem Express Collection	59
3.1.7	InterBioScreen Collection	60
3.1.8	TimTec Collection	60
3.1.9	AK Scientific Collection	60
3.1.10	Natural Products Set IV of the National Cancer Institute's Developmental Therapeutic Program (DTP)	60
3.2	Mixed Collections of Natural Products, Semisynthetic, and Synthetic Compounds	61
4	Coverage and Reach of Molecular Structures Deposited in Natural Product Collections	61
4.1	Coverage of Free and Commercial Virtual Natural Product Collections	62
4.2	Readily Obtainable Natural Products and Derivatives	62
5	Resources for Biological Data on Natural Products	65
6	Resources for Structural Data on Natural Products	65
7	Conclusions	65
	References	66

1 Introduction

Throughout history, natural products have been used as components of traditional medicines and herbal remedies. For modern small-molecule drug development as well, natural products remain the single most productive source of inspiration [1, 2]. According to a widely cited survey of drugs approved between 1981 and 2014 [1], 6% of all small-molecule drugs are unaltered natural products, 26% are natural product derivatives, and 32% are natural product mimetics and/or contain a natural product pharmacophore.

The high importance of natural products is rooted in their evolution-based specific biological purposes, which enable them to exhibit a wide range of biological activities across different organisms. Their structural and physicochemical diversity outrivals that of modern synthetic collections [3–5], and their often high complexity with respect to molecular shape and stereochemistry [3, 6, 7] adds to their ability to modulate a significant number of targets for which no synthetic compounds are known.

Today, in addition to botanicals, natural products from bacteria, fungi, and marine life are increasingly being explored. However, developing drugs from natural products remains a challenging resource- and time-consuming task. Covalent binding, aggregate formation, decomposition, precipitation, and other chemical, physical, and biological processes pose technical barriers to assays run on crude extracts or isolated natural products [2, 8]. Apart from technical complications, the availability of material for testing remains a severe bottleneck. The sourcing process can be complex and expensive, and further complications may arise when material needs to be transferred across national boundaries [2].

Computational methods such as docking, pharmacophore modeling, and quantitative structure–activity relationship modeling can make a significant contribution to natural product-based drug discovery as they allow the selection of promising natural products for extraction, purification, (partial) synthesis, and biological testing [9]. An essential precondition for the application of *in silico* approaches is access to information on the molecular structure of natural products, which today is available from a large number of sources [10]. These sources can be categorized into two main classes: virtual natural product databases and physical natural product collections.

Virtual natural product databases contain the molecular structures of known natural products and vary in size, coverage, and types of information they contain for the individual compounds, among other aspects. As such, they can be further divided into encyclopedic or general, natural product databases, and specialized collections that are focused on, for example, traditional medicines, geographical regions, or bioactivities (e.g., compounds with anticancer or antimalarial activity). The majority of virtual natural product databases are accessible via online services that offer free searching and browsing functionalities. Many of them also offer an option for bulk download, thus enabling virtual screening applications, such as the Dictionary of Natural Products (DNP) [11] and Reaxys [12].

Physical natural product collections are mostly commercial offerings of in-stock natural products and natural products that are sourced or synthesized on-demand. Most vendors make the content of their collections browsable and searchable via free public web services. These web services also often include an option for bulk download. However, the download function may only be enabled after (usually free) registration for the web service.

With this contribution, we aim to provide a timely overview of natural product data sources useful for virtual screening and other applications in cheminformatics. The contribution builds on our recent analyses of virtual natural product databases and physical natural product collections [10, 13] and adds a wealth of information on the latest reported natural product data sources.

2 Virtual Natural Product Databases

In this section, we discuss virtual natural product databases that are particularly relevant for cheminformatics applications in the context of drug discovery. As such, priority is given to resources offering free bulk download of chemical data. At a minimum, the virtual natural product databases listed in this section provide a chemistry-aware web service for browsing and searching, and access to the molecular structures of the search results (Table 1).

Table 1 Overview of virtual natural product databases^a

Data source name	Scope	Number of compounds ^b	Biological data ^c	Free use ^d	Bulk data access	Chemistry-aware web interface	Scientific literature	Web presence and database version	Included in the analysis published in [10]
Encyclopedic and general NP databases									
Dictionary of Natural Products (DNP)	All forms of life	>230,000	Bioactivity data	No	Yes	Yes	–	[11]	Yes
AntiBase	Microorganisms and higher fungi	> 43,000	Bioactivity data (focus on antimicrobial activity)	No	No	Yes	[14]	[15]	No
Reaxys	All forms of life	>260,000	Bioactivity data	No	Yes	Yes	–	[12]	No
Super Natural II	All forms of life	>325,000	Bioactivity and toxicity data	Yes	No	Yes	[16]	[17]	No
UNPD	All forms of life	>229,000	None	Yes	Yes	No	[18]	Website could not be reached	Yes
NPASS	All forms of life	~35,000	Bioactivity data	Yes	No	Yes	[19]	[20]	No
CMAUP	Plants	>47,000	Bioactivity data	Yes	Yes	Yes	[21]	[22]	No
The Natural Products Atlas	Bacteria and fungi	>20,000	None	Yes	Yes	Yes	–	[23]	No
Pye et al. dataset	NPs from microorganisms and marine life published between 2012 and 2015	>6000	None	Yes	Yes	No	[24]	–	No
Natural products included in the PubChem Substance Database	All forms of life	>3500	Bioactivity data	Yes	Yes	Yes	[25]	[26]	Yes

(continued)

Table 1 (continued)

Data source name	Scope	Number of compounds ^b	Biological data ^c	Free use ^d	Bulk data access	Chemistry-aware web interface	Scientific literature	Web presence and database version	Included in the analysis published in [10]
UEFS Natural Products	None specified	~500	None	Via ZINC	Via ZINC	No	-	-	Yes
NP databases focused on traditional medicines									
TCM database@Taiwan	Chinese medicinal herbs	>60,000	Bioactivity data	Yes	Yes	Yes	[27]	[28]	Yes
TCMID 2.0	Chinese medicinal herbs	>43,000	Bioactivity data	Yes	Yes	No	[29]	Website could not be reached	Yes
YaTCM	Chinese medicinal herbs	>47,000	Bioactivity data	Yes	No	Yes	[30]	[31]	No
Chem-TCM	Chinese medicinal herbs	>12,000	Bioactivity data	No	Yes	No	[32]	[33]	No
HIM	Chinese medicinal herbs	~1300	ADME and toxicity data	Yes	Via ZINC	Via ZINC	[34]	Website could not be reached	Yes
HIT	Chinese medicinal herbs	~530	Bioactivity data	Yes	Via ZINC	Via ZINC	[35]	Website could not be reached	Yes
IMPAT	Indian medicinal herbs	>9500	Bioactivity data	Yes	No	Yes	[36]	[37]	No
Databases focused on a specific habitat or geographical region									
DMNP	Marine life	>55,000 (including NP derivatives)	Bioactivity data	No	No	Yes	-	[38]	No

MarinLit	Marine life	>33k	Bioactivity data	No	No	Yes	–	[39]	No
TIPdb	Taiwanese herbs	~9000	Bioactivity data (focus on anticancer, antiplatelet and antituberculosis activity)	Yes	Yes	No	[40, 41]	[42]	Yes
NANPDB	All forms of life indigenous to North Africa	>6800	Bioactivity data	Yes	Yes	Yes	[43]	[44]	Yes
AfroDb	African medicinal plants	~1000	Bioactivity data	Yes	Yes	No	[45]	–	Yes
SANCDDB	South African plants and marine life	>700	None	Yes	Yes	Yes	[46]	[47]	Yes
AfroCancer	African medicinal plants with confirmed antineoplastic, cytotoxic or antiproliferative activity	~400	Bioactivity data (focus on anticancer activity)	Yes	Yes	No	[48]	–	Yes
AfroMalariaDB	African plant NPs with confirmed antimalarial or antiparasitodal activity	>250	Bioactivity data (focus on antimalarial activity)	Yes	Yes	No	[49]	–	Yes
NuBBE _{DB}	NPs from Brazilian plants, fungi, insects, marine organisms, and bacteria	>2200	Bioactivity data (focus on antimicrobial activity)	Yes	Yes	Yes	[50–52]	[53]	Yes
BIOFACQUIM	NPs from plants, fungi, and propolis isolated and characterized in Mexico	>400	Bioactivity data	Yes	Yes	No	[54]	[55]	No
Databases focused on specific organisms									
PAMDB	<i>Pseudomonas aeruginosa</i>	>4300	Bioactivity data	Yes	Yes	Yes	[56]	[57]	No
StreptomeDB 2.0	<i>Streptomyces</i>	~4000	Bioactivity data	Yes	Yes	Yes	[58]	[59]	Yes

(continued)

Table 1 (continued)

Data source name	Scope	Number of compounds ^b	Biological data ^c	Free use ^d	Bulk data access	Chemistry-aware web interface	Scientific literature	Web presence and database version	Included in the analysis published in [10]
Databases focused on specific biological activities									
NPCARE	NPs with measured anti-cancer activity, sourced from plants, marine species and microorganisms	>6500 from online search >1500 in bulk download	Bioactivity data (focus on anticancer activity)	Yes	Yes	No	[60]	[61]	Yes
NPACT	NPs with measured anti-cancer activity, sourced from plants	>1500	Bioactivity data (focus on anticancer activity)	Yes	Via ZINC	Yes	[62]	[63]	Yes
InflamNat	NPs with measured anti-inflammatory activity, sourced primarily from terrestrial plants	>650	Bioactivity data (focus on anti-inflammatory activity)	Yes	Yes	No	[64]	–	No
Databases focused on specific NP classes									
Carotenoids Database	Carotenoids extracted from almost 700 source organisms	>1100	Bioactivity data	Yes	No	Yes	[65]	[66]	No

^aAdapted with permission from [10]. Copyright 2017 American Chemical Society

^bNote that the number of natural products (NPs) stated on websites and provided in data files is often inconsistent, even within the same source. This is because the number of compounds depends, among other aspects, on the exact database version and data parsing and deduplication procedures. Herein are reported what were identified as the most accurate values based on our analysis of original data files, websites, and the primary literature

^cIndicates whether a database includes biological data that can be accessed via a web interface or downloaded

^dIndicates whether the molecular structures of a database are downloadable in bulk or available upon request free of charge

^eIndicates whether a chemistry-aware web interface for browsing and searching (such as exact structure, substructure and similarity search) is provided

2.1 *Encyclopedic and General Natural Product Databases*

2.1.1 Dictionary of Natural Products (DNP)

The Dictionary of Natural Products [11] is one of the most established encyclopedic collections of natural products available to date. The commercial database consists of more than 230k natural products, 46k of which are not covered by any of the free virtual natural product collections investigated in our recent study [10] and marked in Table 1. The molecular structures are richly annotated with compound names and synonyms, physicochemical properties (e.g., molecular weight, p*K*_a, solubilities, and spectroscopic data), biological sources, use, and toxicity data. One particularly useful feature of this database is that the natural products are classified into 1050 structural types. Importantly, stereochemical information is stored only in Fisher-type diagrams, separate from the 2D connection tables and InChIs. The database is accessible via a web service [11] and also distributed as a CD-ROM.

2.1.2 AntiBase

AntiBase [15] is a comprehensive commercial database including more than 43k natural products collected primarily from microorganisms and higher fungi (including algae, cyanobacteria, lichens, yeasts, Ascomycetes, and Basidiomycetes). AntiBase stands out due to the large amount of spectrometric data provided (including experimental and computed ¹³C NMR data). The individual natural products are annotated with further physicochemical properties and biological data, such as pharmacological activities and toxicity. AntiBase is available in several software formats featuring powerful text, structure, and spectra search capabilities.

2.1.3 Reaxys

Reaxys [12] is a comprehensive resource for chemical information relevant to synthesis chemists. As such, Reaxys has no specific focus on natural products, but contains information on the molecular structures, reactions, physical properties, biological sources, and activity data for more than 260,000 natural products. Reaxys is accessible via a web interface, which features detailed search functionality. Bulk download of natural products (and other chemicals and data) is supported.

2.1.4 Super Natural II

Super Natural II [16] provides chemical information on more than 325,000 natural products and, accordingly, is currently one of the most comprehensive free data sources available. Super Natural II draws data from several preexisting databases

and provides information on molecular structures (including stereochemistry annotations), suppliers, bioactivities, computed physicochemical properties, and toxicity classes. The web interface supports the download of individual structures but not bulk download.

2.1.5 Universal Natural Products Database (UNPD)

With a total of more than 229,000 entries, the Universal Natural Products Database (UNPD) [18] is currently the most comprehensive of all free and commercial resources on natural products that offer bulk download. Drawing data from a number of different sources, including the Chinese Natural Product Database (CNPD) [67], the CHDD [68] (a database of compounds of traditional Chinese medicinal herbs, previously provided by the authors of the UNPD), and the Traditional Chinese Medicines Database (TCMD) [69], the UNPD is itself a component of Super Natural II. Our recent analysis showed that approximately one-third of the natural products contained in the UNPD are not covered by any of the other investigated virtual natural product databases [13]. We also found that the UNPD covers a wide chemical space and represents all major classes of natural products. Approximately 85% of the natural products contained in the UNPD comply with Lipinski's rule of five (here and elsewhere, statements on the compliance with Lipinski's rule of five refer to the molecular structures of natural products after the removal of sugars and sugar-like moieties with the tool "SugarBuster" [13]). The connection tables of UNPD store 3D structures with explicit stereochemistry defined by atom coordinates (enantiomers are stored as individual entries) plus several identifiers. In recent years, significant downtimes of the web presence have been observed.

2.1.6 Natural Product Activity and Species Source (NPASS)

The Natural Product Activity and Species Source [19] is another large resource of chemical and biological information on natural products. The database currently includes more than 35,000 natural products from a total of approximately 25,000 species. Two-thirds of the natural products come from Viridiplantae; the remaining third comes primarily from Metazoa, fungi, and bacteria. Bioactivity data are recorded against approximately 3000 protein targets, more than 1300 microbial species and a similar number of cell lines. Natural Product Activity and Species Source offers a powerful, chemistry-aware web interface for browsing and searching. Data for individual natural products can easily be downloaded, but bulk download of structures and other data is not offered.

2.1.7 Collective Molecular Activities of Useful Plants (CMAUP)

Collective Molecular Activities of Useful Plants [21] is a large, new resource for information on plant natural products and their biological activities. The database stores information on over 47,000 natural products of more than 5600 plants native to greater than 150 countries and regions. The individual natural products are annotated with recorded bioactivities against more than 640 biomacromolecular targets. In addition, information on plant species, use, geographical distribution, metabolic pathways, gene ontologies, and diseases is provided. The database can be browsed and searched via a free, chemistry-aware web interface. Free bulk download of structural data (including stereochemical information) and metadata is also supported.

2.1.8 Natural Product Atlas

The Natural Product Atlas [23] has been recently introduced as a comprehensive resource of chemical information on natural products from bacteria (including cyanobacteria) and fungi (including mushrooms and lichens) reported in peer-reviewed original research articles. The current version of the database covers approximately 20,000 natural products, almost one-third of which are found in *Streptomyces*. Further prominent genera are *Aspergillus* and *Penicillium*, each representing approximately 10% of the data. The web service provides powerful tools for browsing, searching, and data visualization. Particularly noteworthy are the network visualization features, which allow users to obtain a solid overview of the molecular diversity and coverage of the chemical space. An option for bulk download of the database is provided.

2.1.9 Pye et al. Dataset

As part of a comprehensive survey of natural products discovered between 1941 and 2015, Pye et al. have recently published a dataset of almost 6300 natural products that have been published between 2012 and 2015 [24]. As such, the dataset provides a good overview of the chemical space of natural products discovered in recent years. All structures are available as isomeric SMILES (simplified molecular input line entry specification) from the supporting information.

2.1.10 Natural Products Included in the PubChem Substance Database

The PubChem database [70] contains structures of more than 3500 natural products, which can be retrieved using the query “MLSMR [SRC] AND NP[CMT]” [25]. Most compounds are annotated with bioactivity data, covering a total of

more than 650 biomolecular targets. Approximately 40% of all compounds are not covered by any other resource investigated in our recent study [13]. More than 95% of all natural products of this dataset comply with Lipinski's rule of five; greater than half of all compounds are alkaloids. All structures are downloadable and include stereochemical information.

2.1.11 UEFS Natural Products

Researchers from the State University of Feira de Santana (UEFS) in Brazil have deposited a dataset of approximately 500 natural products for download at the ZINC database [71, 72]. The natural products have been compiled from papers that the authors and collaborators have published separately. Noteworthy is the relatively high proportion of flavonoids in the dataset [13].

2.2 Databases Focused on Traditional Medicines

2.2.1 Traditional Chinese Medicine Database@Taiwan

The TCM Database@Taiwan [27] is the most comprehensive free resource for molecular structures of natural products related to TCM. It has been compiled from Chinese medical texts and various dictionaries, and contains the structures of more than 60,000 natural products from over 450 herb, animal, and mineral product TCMs. Important features of this database include the organization of the data into 22 TCM usage classes, such as "digestant medicinal", and comprehensive ingredient-to-TCM mapping. We found that 38% of all natural products of the TCM Database@Taiwan are alkaloids, which is one of the highest percentages observed among all investigated databases [13]. The database also stands out due to its large proportion of high molecular weight natural products, among which polyphenols and basic alkaloids are particularly prominent. In contrast to the previously discussed natural product databases, the proportion of natural products in compliance with Lipinski's rule of five is only 51%. The web interface of the TCM Database@Taiwan offers advanced search functionalities based on molecular structures and physicochemical properties. Bulk download of all molecular structures including stereochemical information is supported.

2.2.2 Traditional Chinese Medicine Integrated Database (TCMID 2.0)

The TCMID 2.0 [29] is a large database of natural products that links traditional Chinese with modern western medicine by incorporating data on drugs, targets, and diseases. The database integrates data on herbal ingredients from, among many other sources, the TCM Database@Taiwan, TCM-ID [73], and the Encyclopedia of

Traditional Chinese Medicines [74]. Since its initial release in 2013, the database has been substantially expanded, with the latest release counting more than 43k compounds. As major additions to the latest release, almost 4k mass spectra of natural products and over 176,000 protein-protein interactions have been added. The TCMID 2.0 web interface offers, among many other features, a tool for visualizing ingredient-target-drug-disease networks and herb-target-disease networks. This enables users, for example, to browse the natural products of a herb of interest, the targets of these natural products and how they are linked to diseases. As such, the platform can provide valuable information on multi-target effects and molecular mechanisms. Download of molecular structures (including stereochemical information) and associated data is possible in principle. At the time of writing, the online presence of this database could not be confirmed.

2.2.3 Yet Another Traditional Chinese Medicine Database (YaTCM)

The YaTCM database [30] is a further recently introduced database on natural products from Chinese medicinal herbs. The database currently holds more than 47,000 records of natural products found in over 6200 herbs. Like TCMID 2.0 (which is integrated into YaTCM), the chemical data are supplemented with a wealth of information on targets (approximately 3500 therapeutic targets are covered), pathways, and diseases. The web service offers chemistry-aware browsing and search functionality. The website also features an *in silico* model for target prediction and tools for visualizing networks of TCM recipes, herbs, natural products, known and predicted protein targets, pathways, and diseases. Bulk download of chemical information is not supported.

2.2.4 Chemical Database of Traditional Chinese Medicine (Chem-TCM)

Chem-TCM [33] is a commercial resource that holds more than 12,000 records on natural products from approximately 350 herbs used in TCM. The database provides rich chemical information, including molecular structures with stereochemical information, names and identifiers, molecular scaffold types, and natural product classes. The botanical information includes Latin binomial botanical names, pharmaceutical names, and Chinese herb names. Chem-TCM seeks to link TCM to western medicine by including activities against 41 drug targets predicted with a random forest model [32]. In addition, the database includes estimated affinities of molecular activities according to 28 traditional Chinese herbal medicine categories. Chem-TCM is provided via a chemistry-aware software application and as SD files.

2.2.5 Herbal Ingredients In Vivo Metabolism Database (HIM)

The Herbal Ingredients In Vivo Metabolism (HIM) [34] consists of around 1300 natural products richly annotated with absorption, distribution, metabolism, and excretion (ADME) data and information on compound toxicity. Most natural products of HIM comply with Lipinski's rule of five, and approximately one-third of the natural products in this database are not available from any of the other resources that we investigated recently [13].

At the time of writing, the online presence of this database could not be confirmed. The molecular structures of HIM can, however, be accessed via the ZINC database and include stereochemical information.

2.2.6 Herbal Ingredients' Targets Database (HIT)

The Herbal Ingredients' Targets (HIT) database [35] is a collection of more than 530 active ingredients from herbs. Most natural products of HIT comply with Lipinski's rule of five [13]. As for HIM, the web presence of HIT could not be confirmed at the time of writing, but the molecular structures (including stereochemical information) are available via the ZINC database. The natural products stored in HIT are covered to a large extent by other databases [13].

2.2.7 Indian Medicinal Plants, Phytochemistry, and Therapeutics Database (IMPPAT)

The Indian Medicinal Plants, Phytochemistry, and Therapeutics (IMPPAT) database [36] is a rich resource of chemical, biological, and botanical information on Indian medicinal plants, covering more than 9500 natural products from more than 1700 species. The chemistry-aware web interface allows browsing and searching. A network visualization tool allows the investigation of plant-natural product associations, plant-therapeutic use associations, and plant-formulation associations. Bulk download of molecular structures is not supported.

2.3 Databases Focused on a Specific Habitat or Geographic Region

2.3.1 Dictionary of Marine Natural Products (DMNP)

The Dictionary of Marine Natural Products [38] is a subset of the Dictionary of Natural Products (DNP) containing more than 55,000 marine natural products and their derivatives. This commercial resource is provided as a web service (with

similar capacities as that of the DNP) and is also distributed as a combination of a book and CD-ROM.

2.3.2 MarinLit Database

The MarinLit database [39] is a large database of marine natural products collected from journal articles. The commercial resource currently lists more than 33,000 natural products, richly annotated with bibliographic information, molecular structure, names, biological sources, physicochemical properties, and identifiers. MarinLit's web interface provides powerful search functionalities and features for the dereplication of natural products.

2.3.3 Taiwan Indigenous Plant Database (TIPdb)

The TIPdb database [40] provides information on the anticancer, antituberculosis, and antiplatelet activity of more than 9000 natural products of plants indigenous to Taiwan. Noteworthy are the rather high percentage of natural products with sugars and sugar-like moieties (25%) and a rather low percentage of alkaloids (14%) [13]. The web service offers basic browsing and searching functionality, and the molecular structures of all natural products can be downloaded in bulk.

2.3.4 Northern African Natural Products Database (NANPDB)

With more than 6800 natural products records, NANPDB [43] is the largest database of natural products isolated from species native to Northern Africa, primarily plants but also endophytes, animals, fungi, and bacteria. This freely accessible database has been compiled from many different sources, including articles published in natural product journals as well as Ph.D. theses. The database provides information on source organisms, biological activities, and activity types (e.g., antimalarial, cancer-related). We have shown that the chemical space covered by NANPDB is similar to that of approved drugs, with more than 90% of all compounds complying with Lipinski's rule of five [13]. Noteworthy is the high proportion of natural products containing sugars and sugar-like moieties (28%). The Northern African Natural Products Database is provided via a chemistry-aware web interface [44] and can be downloaded in SMILES and SD file format (including stereochemical information).

2.3.5 AfroDb Database

The AfroDb database [45] is a diverse collection of natural products found in African medicinal plants. Worth mentioning is the high percentage of phenols and phenol

ethers in this database (61%), which is approximately double of that of the DNP [13]. The molecular structures (including stereochemical information) are freely available in the supplementary information of the original publication and via the ZINC database.

2.3.6 South African Natural Compound Database (SANCDB)

The SANCDB [46] is composed of more than 700 natural products from plants and marine life native to South Africa. The database has been compiled manually from the literature and contains information on molecular structure (including stereochemistry information), name, structural class, source organism, and physicochemical properties. A free, chemistry-aware web interface for searching and browsing is provided. The resource is also accessible via a representational state transfer application programming interface (REST API).

2.3.7 African Anticancer Natural Products Library (AfroCancer)

AfroCancer [48] focuses on natural products from African medicinal plants with confirmed antineoplastic, cytotoxic, or antiproliferative activity. The database contains a high percentage of phenols and phenolic compounds (57%) [13]. The molecular structures (including stereochemical information) are freely available in the supplementary information of the original publication.

2.3.8 African Antimalarial Natural Products Library (AfroMalariaDB)

The AfroMalariaDB [49] is focused on natural products with antimalarial or antiplasmodial activity confirmed by *in vitro* and/or *in vivo* experiments. It consists of approximately 250 natural products collected from more than 130 African plants. Like AfroDb and AfroCancer, AfroMalariaDB is rich in phenols and phenolic compounds [13]. The database is available for download in the supplementary information of the original publication.

2.3.9 Nuclei of Bioassays, Biosynthesis, and Ecophysiology of Natural Products Database (NuBBEDB)

The NuBBE database [50, 51] lists more than 2200 natural products of mainly plants but also fungi, insects, marine organisms, and bacteria native to Brazil. In addition to chemical information, pharmacological and toxicological data are provided. Most of the natural products contained in NuBBEDB are drug-like [50]. Compared to other sources, a low proportion of alkaloids (9%) is observed [50]. The chemistry-aware web interface allows the search for compounds according to structure, spectroscopic

information, physicochemical properties, and biological source. Bulk download of structures in MOL2 file format is available.

2.3.10 BIOFACQUIM Database

The BIOFACQUIM database [54] is a manually compiled dataset of natural products isolated and characterized in Mexico. Approximately three-quarters of the 400 natural products currently listed in this database are from plants and 23% are from fungi. The web service offers basic searching functionality and bulk download of all data (molecular structures including stereochemical information).

2.4 Databases Focused on Specific Organisms

2.4.1 *Pseudomonas aeruginosa* Metabolome Database (PAMDB)

The PAMDB [56] is a rich resource of natural products found in *Pseudomonas aeruginosa*. The database contains more than 4300 natural products linked to ontology, reaction, and pathway data. The database also provides information on the physicochemical properties of natural products and cross-links to external resources. The PAMDB can be browsed and searched via a chemistry-aware web interface [57]. The web service also offers bulk download of data in various formats.

2.4.2 StreptomeDB 2.0

StreptomeDB 2.0 [58] is a comprehensive database of about 4000 natural products produced by Streptomyces. The database has been compiled from the literature, the Novel Antibiotics Database [75], and KNApSAcK [76, 77]. The individual molecular structures (including stereochemical information) are annotated with names, *Streptomyces* species, biological activities, and key physicochemical properties. Approximately one-third of the natural products recorded in StreptomeDB2.0 are not available from any of the other resources that we investigated recently [13]. StreptomeDB2.0 stands out by having one of the largest proportions of natural products containing sugars and sugar-like moieties (25%). Although most of the natural products of StreptomeDB2.0 cover areas in chemical space that are also densely populated with approved drugs, only a relatively small portion of the natural products in this database comply with Lipinski's rule of five (70%). Noteworthy are a high proportion of alkaloids (47%), although only relatively few of these contain a basic nitrogen (19%). The database can be freely searched and browsed via a chemistry-aware web interface. Bulk download of the data in SD file format with chirality flags is supported.

2.5 Databases Focused on Specific Biological Activities

2.5.1 Database of Natural Products for Cancer Gene Regulation (NPCARE)

The NPCARE database [60] contains more than 6500 natural products with potential anticancer activity measured for a total of approximately 1100 cell lines for 34 cancer types. The natural products in NPCARE originate from more than 2000 plants, marine species, and microorganisms. The provided data include chemical information (including molecular structures with stereochemistry annotations) and information on modulated genes and proteins. The molecular structures of a subset of more than 1500 compounds are available for bulk download (the SMILES notations do not include stereochemical information; however, this information can be retrieved using the PubChem compound identifiers provided).

2.5.2 Naturally Occurring Plant-Based Anti-cancer Compound-Activity-Target Database (NPACT)

The NPACT database [62] is focused on plant-derived natural products with experimentally confirmed cancer-inhibitory activity. The database lists more than 1500 compounds annotated with approximately 5200 compound-cell line and 2000 compound-target interactions. Cross-links with other resources such as the HIT database and PubChem are also provided. The chemistry-aware web interface allows browsing and searching. The molecular structures including stereochemical information can be downloaded from the ZINC database.

2.5.3 InflammNat Database

The InflammNat database [64] contains 665 natural products with experimentally confirmed anti-inflammatory activity. Most natural products (86%) originate from terrestrial plants; a minority comes from marine life, terrestrial fungi, and bacteria. The InflammNat database is rich in flavonoids and triterpenoids. Cross-linking with the PubChem Bioassay database provides information on the biomolecular targets of the natural products. All structures are provided in the supporting information of the publication on InflammNat.

2.6 Databases Focused on Specific Natural Product Classes

2.6.1 Carotenoids Database

The Carotenoids Database [65] contains over 1100 natural carotenoids extracted from almost 700 source organisms. The resource was compiled from the primary literature. The web interface provides access to molecular structures, source organisms, and biological function of the individual carotenoids. The structures of individual carotenoids can be downloaded in various formats (including stereochemical information) but only one molecule at a time.

3 Physical Natural Product Collections

Few physical collections are in existence that are purely based on genuine natural products. More common are physical collections containing a mix of natural products, natural product analogs and derivatives, and synthetic compounds. Among the mixed collections, only a minority have annotated their compounds as genuine natural products, semisynthetic, and synthetic compounds. However, computational approaches allow the accurate discrimination of natural products and (semi-)synthetic compounds based on molecular structures. The latest *in silico* approach, “NP-Scout”, has been reported from our lab [78]. The NP-Scout approach is a random forest-based machine-learning model that calculates the probability of a compound being a natural product. The model was trained on more than 265,000 natural products and synthetic molecules. On an independent test set of over 80,000 compounds, the model reached an area under the receiver operating characteristic curve (AUC) of 0.997 and a Matthew’s correlation coefficient (MCC) of 0.960, documenting the high performance of the model. The NP-Scout web service also supports the generation of similarity maps, which indicate atoms in a molecule that contribute significantly to the classification of a molecule as a synthetic molecule or natural product. This allows, for example, the identification of synthetic fragments in natural product derivatives. Two examples of similarity maps generated with NP-Scout are shown in Fig. 1, for vorapaxar and empagliflozin. Vorapaxar is a derivative of the natural product himbacine, for which NP-Scout correctly identifies the decahydronaphtho[2,3-*c*]furan-1(3*H*)-one scaffold as being a natural product fragment. Empagliflozin mimics the flavonoid phlorizin, and NP-Scout correctly recognizes the *C*-glycosyl moiety as being a natural product fragment.

In the following Sections, we will discuss examples of physical natural product collections for which molecular structures are accessible via a chemistry-aware web interface and/or bulk download. An overview of the resources discussed herein is provided in Table 2.

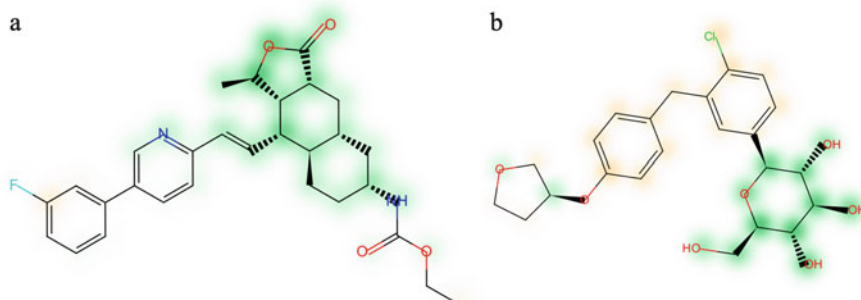


Fig. 1 Similarity maps of (a) vorapaxar and (b) empagliflozin. Green-highlighted atoms contribute to the classification of a molecule as a natural product; orange-highlighted atoms contribute to the classification of a molecule as a synthetic compound. Adapted from [78] (CC BY 4.0; <https://creativecommons.org/licenses/by/4.0>)

Table 2 Physical natural product collections^a

Supplier name	(Sub-)set name	Number of compounds	Collection composition	Molecular structures provided free of charge	Web presence
Ambinter and Greenpharma	Natural products	>8000; plated collection of 480 NPs	NPs only	Yes	[79, 80]
Ambinter and Greenpharma	Natural product derivatives	>11,000	(Semi-) synthetic compounds	Yes	[79, 80]
AnalytiCon Discovery	MEGx—Purified natural products of microbial and plant origin	~5000	NPs only	Yes	[81]
AnalytiCon Discovery	NATx—Semi-synthetic natural product-derived compounds	>29,000	NPs and (semi-) synthetic compounds	Yes	[81]
AnalytiCon Discovery	MACROx—Next generation macrocycles	>2000	Semisynthetic compounds based on nine scaffolds	Yes	[81]
AnalytiCon Discovery	FRGx—Fragments from Nature	>200	NPs and (semi-) synthetic compounds	Yes	[81]
Chengdu Biopurify Phytochemicals	TCM Compounds Library	>4600	NPs and (semi-) synthetic compounds	Yes	[82]

(continued)

Table 2 (continued)

Supplier name	(Sub-)set name	Number of compounds	Collection composition	Molecular structures provided free of charge	Web presence
Selleck Chemicals	Natural Products	~1600 (plated)	NPs only	Yes	[83]
TargetMol	Natural Compound Library	>1500 (plated)	NPs only	Yes	[84]
MedChem Express	Natural Product Library	>1500; plated collection of >900 NPs	NPs only	Yes	[85]
InterBioScreen	Natural Compound (NC) Collection	>1300 natural compounds and 66,000 derivatives and analogs	NPs and (semi-) synthetic compounds; distinguishable by tags	Yes	[86]
InterBioScreen	Building Blocks	>13,000	NPs and (semi-) synthetic compounds	Yes	[86]
InterBioScreen	Natural Scaffold Libraries	>500	NPs and (semi-) synthetic compounds	Yes	[86]
TimTec	Natural Product Library (NPL)	~800	NPs only	No	[87]
TimTec	Natural Derivatives Library (NDL)	~3000	NPs and (semi-) synthetic compounds	Yes	[87]
TimTec	Flavonoids Collection	~500	NPs and (semi-) synthetic compounds	Yes	[87]
TimTec	Flavonoid Derivatives Extended Collection	>4000	NPs and (semi-) synthetic compounds	Yes	[87]
TimTec	Gossypol Derivatives Collection	~100	NPs and (semi-) synthetic compounds	Yes	[87]
AK Scientific	Natural Products	~500	NPs only	Yes	[88]
Developmental Therapeutic Program (DTP) of NCI NIH	Natural Products Set IV	~400	NPs only	Yes	[89]
INDOFINE Chemical	Natural Products, Flavonoids, Coumarins, etc.	>4000	NPs and (semi-) synthetic compounds	Yes	[90]

(continued)

Table 2 (continued)

Supplier name	(Sub-)set name	Number of compounds	Collection composition	Molecular structures provided free of charge	Web presence
Pharmeks	Screening Compounds	>360,000 (>2800 NPs and NP derivatives)	NPs and (semi-) synthetic compounds; distinguishable by tags	Yes	[91]
Pharmeks	Building Blocks	>12,000	NPs and (semi-) synthetic compounds	Yes	[91]
Princeton Bio-Molecular Research	Macrocycles	>1500	NPs and (semi-) synthetic compounds	Yes	[92]
MicroSource Discovery Systems	Natural Products Collection (NatProd)	~800	NPs and (semi-) synthetic compounds	Yes	[93]
Specs	Natural Products	>600	NPs and (semi-) synthetic compounds	Yes	[94]

^aAdapted with permission from [10]. Copyright 2017 American Chemical Society

3.1 Pure Natural Product Collections

In this section, we list offerings of pure natural product collections and mixed collections in which genuine natural products are clearly marked and can hence be distinguished from other compounds.

3.1.1 Ambinter and Greenpharma

With more than 8000 listed compounds, the physical natural product collection of Ambinter and Greenpharma [79] is one of the largest offerings available to date. As we have shown previously [13], approximately half of all these natural products are available exclusively from these providers. The collection stands out due to the well-balanced representation of all major natural product classes, which is comparable to that observed for the DNP [13]. Ambinter and Greenpharma also offer a collection of more than 11,000 purchasable natural product derivatives and a preformatted collection of 480 diverse natural products.

3.1.2 AnalytiCon Discovery

AnalytiCon Discovery [81] offers a continuously growing collection of purchasable natural products (“MEGx”). The collection consists of approximately 5000 compounds, the majority of which are available exclusively from this provider [13]. Among the offered compounds are many microbial natural products. The MEGx has the highest proportion of natural products containing sugar or sugar-like fragments among all natural product collections we investigated previously. In contrast, the percentage of alkaloids in this collection is low (14%). AnalytiCon also offers collections of more than 29,000 semisynthetic compounds derived from natural products (“NATx”), over 2000 macrocycles (“MACROx”), and more than 200 fragments from Nature (“FRGx”).

3.1.3 Chengdu Biopurify Phytochemicals

Chengdu Biopurify Phytochemicals [82] offers a collection of over 4600 compounds related to TCM. The collection is rich in flavonoids, alkaloids, phenols, and terpenoids. Many of the natural products are offered exclusively by this provider.

3.1.4 Selleck Chemicals

Selleck Chemicals [83] offers a plated collection of over 1600 natural products for screening. The collection is rich in flavonoids and phenolic natural products, and more than three-quarters of the natural products in this collection comply with Lipinski’s rule of five [13].

3.1.5 TargetMol Collection

TargetMol [84] offers a plated collection of more than 1500 natural products for screening. The compounds originate from plants, animals, microorganisms, and other organisms. Many of the natural products of this collection are active on pharmaceutically relevant proteins.

3.1.6 MedChem Express Collection

The MedChem Express collection [85] offers a diverse ensemble of more than 1500 natural products, including 216 alkaloids, 189 terpenoids and glycosides, 183 acids and aldehydes, 156 flavonoids, and 88 saccharides and glycosides. The company also offers a plated collection of more than 900 natural products for screening.

3.1.7 InterBioScreen Collection

InterBioScreen [86] offers the Natural Compound (NC) collection of purchasable compounds, which contains over 1300 genuine natural products plus 66,000 natural product derivatives (the labels allow the discrimination of genuine natural products from natural product analogs and derivatives). The vast majority of natural products contained in this collection originate from plants, 5 to 10% are isolated from microbes, and another 5% from marine species. The NC collection includes uncommon compounds as well, such as certain classes of phytoalexins, allelopathic agents, and specific sex attractants. In our recent studies, we found that the NC collection features the highest rate of steroids among all investigated natural product databases [13]. Approximately 95% of all compounds of the natural product collection comply with Lipinski's rule of five. InterBioScreen also offers a collection of over 13,000 building blocks that are partly related to natural products, plus more than 500 natural product scaffolds for compound synthesis.

3.1.8 TimTec Collection

The Natural Product Library (NPL) from TimTec [87] consists of approximately 800 genuine natural products. These natural products originate primarily from plants, but some have animal, bacterial, or fungal origins. In addition, TimTec offers the Natural Derivatives Library (NDL), which is composed of more than 3000 natural product derivatives, natural product analogs, and semi-natural compounds. A subset of 500 flavonoid derivatives based on nine core flavonoid scaffolds is available, as are an extended collection of over 4000 flavonoid derivatives and a small collection of gossypol derivatives.

3.1.9 AK Scientific Collection

AK Scientific [88] offers a collection of approximately 500 natural products including alkaloids, flavonoids, stilbenoids, terpenoids, and terpenes. The company also provides a subset of synthetic compounds and additives, containing over 100 flavonoids, food preservatives/additives, and vitamins.

3.1.10 Natural Products Set IV of the National Cancer Institute's Developmental Therapeutic Program (DTP)

The Developmental Therapeutic Program of the National Cancer Institute, National Institutes of Health, provides a plated collection of approximately 400 natural products for experimental screening. These natural products have been selected from 140,000 compounds available from the DTP Open Repository based on compound diversity, availability, and purity. According to our previous analysis

[13], more than 60% of these compounds are available exclusively from this source. Approximately 80% comply with Lipinski's rule of five, which is the lowest among all investigated physical collections. Noteworthy is the high proportion of alkaloids (42%).

3.2 Mixed Collections of Natural Products, Semisynthetic, and Synthetic Compounds

More than 100 vendors offer natural products for experimental testing today, as will be discussed in the next section. However, only a rather small number of vendors explicitly mention the presence of natural products in their mixed physical collections. One of them is INDOFINE Chemical [90], which offers around 4000 natural products and semisynthetic compounds including flavones, isoflavones, flavanones, coumarins, chromones, chalcones, and lipids. The company also has a broad portfolio of synthetic compounds.

Pharmeks [91] offers a diverse, mostly heterocyclic collection of 360,000 organic molecules, 2800 of which are natural products or natural product derivatives. In addition, Pharmeks also offers more than 12,000 building blocks of both synthetic compounds and natural products.

Princeton BioMolecular Research [92] provides a collection of over 1500 macrocyclic natural products, natural product derivatives, and synthetic compounds. MicroSource Discovery Systems [93] offers its Natural Products Collection ("NatProd"), which is composed of 800 natural products and natural product derivatives originating from plant, animal, and microbial sources. Specs [94] offers a collection of over 600 isolated or synthesized natural products and natural product derivatives originating from fungi, bacteria, plants, marine species, and other organisms.

4 Coverage and Reach of Molecular Structures Deposited in Natural Product Collections

As part of one of our previous studies [10], we have analyzed the coverage and reach of 18 virtual natural product databases (marked in Table 1 as included in the analysis published in Ref. [10]) and several physical natural product collections. The number of unique compounds contained in the individual datasets was determined by counting unique InChIs (without stereochemistry and fixed hydrogen layers) derived from neutralized molecules (i.e., counter-ions of salts removed and compounds neutralized with the Wash function in the Molecular Operating Environment (MOE) [95]). Summarized here are some of the most relevant findings of this study.

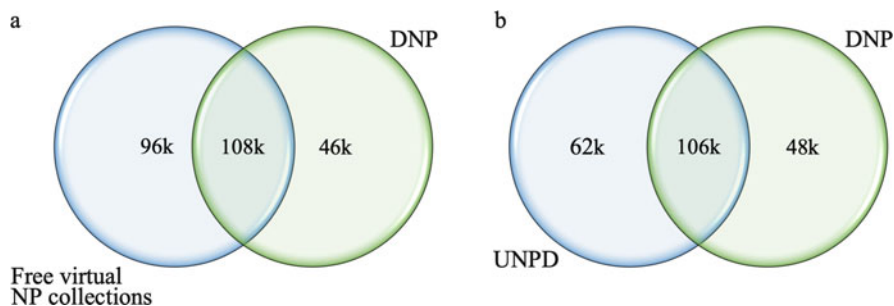


Fig. 2 The overlap between the Dictionary of Natural Products (DNP) and (a) the freely accessible virtual natural product collections or (b) the Universal Natural Products Database (UNPD). Reprinted with permission from [10]. Copyright 2017 American Chemical Society

4.1 Coverage of Free and Commercial Virtual Natural Product Collections

The 18 virtual natural product databases marked in Table 1 contain more than 250,000 unique natural products in total. Approximately 46,000 of these natural products are exclusively covered by the DNP, which is the most widely accepted reference natural product database (Fig. 2a). At the same time, 70% of all natural products listed in the commercial DNP are also present in at least one free database. The largest contribution to the significant overlap between the DNP and the free virtual natural product collections stems from the UNPD, which remains the most comprehensive free and fully downloadable virtual natural product database.

4.2 Readily Obtainable Natural Products and Derivatives

In the context of early drug discovery, virtual screening in particular, it is important to understand both the proportion of and coverage of chemical space by natural products that are readily obtainable for experimental testing. Only approximately 11,000 natural products are readily obtainable from pure, physical natural product collections. However, the number increases to more than 25,000 when also taking mixed physical collections into account. This number was derived by overlaying a dataset of 250,000 known natural products (sources marked in Table 1) with the 7.3 million readily obtainable compounds listed in the ZINC database “in-stock” subset (Fig. 3). The ZINC database is widely accepted as the most comprehensive meta-database of purchasable compounds and offers a subset of readily obtainable compounds. As part of this analysis, 100 vendors of natural products were identified. Only nine of these offer more than 5000 readily obtainable compounds (Table 3). The number of accessible natural products can be further increased by using services for on-demand sourcing, extraction, and synthesis. This involves longer lead times

Fig. 3 Comparison of the content of virtual natural product collections and the ZINC “in-stock” subset. Reprinted with permission from [10]. Copyright 2017 American Chemical Society

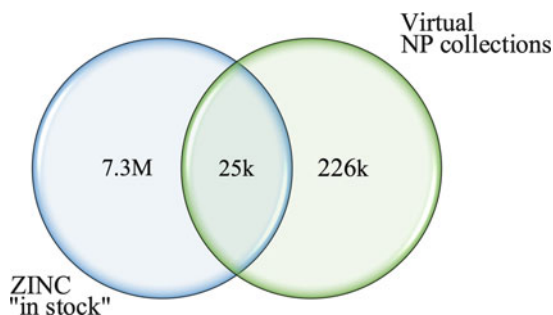


Table 3 Numbers of natural products readily purchasable from suppliers^a

Number of readily purchasable NPs	Suppliers
>5000	Molport, TimTec, AK Scientific, Tetrahedron Scientific, BOC Sciences, FineTech Industry, Sigma Aldrich, Specs, National Cancer Institute (NCI)
3000 to 5000	Fluorochem, Nanjing Kaimubo Pharmatech Company, Hong Kong Chemhere, Oxchem Corporation, BePharm, Zelinsky Institute, Combi-Blocks, Debye Scientific, Matrix Scientific, WuXi AppTec, Ark Pharm, Bide Pharmatech, BioSynth, InterBioScreen, Labseeker, StruChem, Alfa-Aesar
2000 to 3000	AstaTech, Enamine, Oakwood Chemical, Frontier Scientific Services, Alfa Chemistry, Key Organics, Apollo Scientific, W&J PharmaChem, AnalytiCon Discovery, Acros Organics, Shanghai Pi Chemicals, Syntharise Chemical
1000 to 2000	Toronto Research Chemicals, Capot Chemical, Rostar, INDOFINE Chemical, Alinda, Pharmeks, Innovapharm, Synthon-Lab, Vesino Industrial, Life Chemicals, Bosche Scientific, Chem-Impex International, Vitas-M Laboratory, Biopurify Phytochemicals, Otava Chemicals, A2Z Synthesis, Cayman Chemical, Accela ChemBio, Molepedia, Curpys Chemicals, ChemDiv, AsisChem
100 to 1000	Boerchem Pharmatech, AbovChem, Ryan Scientific, Hangzhou Yuhao Chemical Technology, TargetMol, APExBIO, Princeton BioMolecular Research, EDASA Scientific, ChemBridge, Maybridge, MolMall, HDH Pharma, UORSY, Chemik, Bachem, Creative Peptides, MedChem Express, Aronis, Heteroz, Selleck Chemicals, Tocris, Frinton Laboratories, Asinex, Synchem, EndoTherm Life Science Molecules, Coresyn, SpiroChem, Advanced ChemBlock

^aNumbers are estimates based on the overlap of all known natural products (NPs) and the compounds present from a particular vendor in the “in-stock” subset of ZINC. Reprinted with permission from [10]. Copyright 2017 American Chemical Society

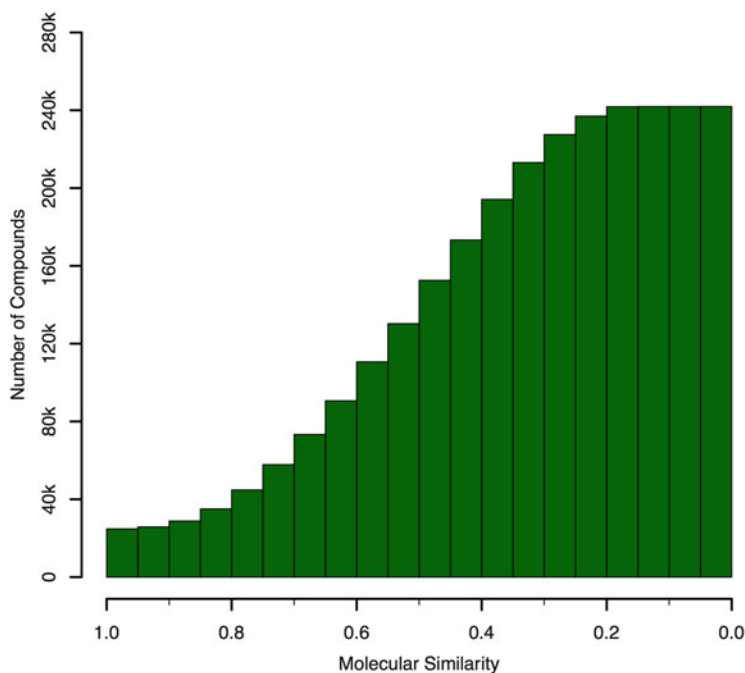


Fig. 4 Cumulative histogram of maximum molecular similarity (Tanimoto coefficient) for the compounds in virtual natural product libraries compared to the ZINC “in-stock” subset. The bars in the histogram represent the number of known natural products with a maximum molecular similarity greater than or equal to the bin threshold. Reprinted with permission from [10]. Copyright 2017 American Chemical Society

and higher costs but, as Lucas et al. [96] have shown recently, approximately one-third of all natural products listed in the DNP, TCM Database@Taiwan, and StreptomeDB are obtainable via these routes.

As observed in the physical collection sizes reported in Table 2, the number of readily obtainable natural product analogs and derivatives is much higher than that of genuine natural products. Hence, by allowing small deviations in molecular structure from genuine natural products, a much higher number of natural product-like compounds become readily obtainable. As shown in Fig. 4, there are approximately 58,000 natural products readily obtainable that have a Tanimoto coefficient based on Morgan3 fingerprints [97] equal to 0.7 or higher. Given these high similarity values, these compounds are likely natural product derivatives or analogs.

Macrocycles have gained significant interest in the context of drug discovery in recent years. Due to their conformational constraints, macrocycles can provide advantages in entropic binding and specificity [98]. Our analysis has shown that approximately 14% (35,000) of all 250,000 known natural products contain rings formed by more than seven atoms. However, only approximately 800 genuine natural products with a ring size larger than seven atoms are readily obtainable

(note that, e.g., AnalytiCon offers more than 2000 semisynthetic, macrocyclic compounds based on nine scaffolds).

5 Resources for Biological Data on Natural Products

The majority of virtual natural product databases provide biological information in addition to chemical data (Table 1). Most of this information is in the form of bioactivities measured for organisms, cells, or individual biomacromolecules. Several resources provide information on pathways, diseases, and ADME properties.

The ChEMBL [99] database is one of the most comprehensive sources of measured biological activities of small molecules. The database is manually compiled primarily from scientific publications. It also draws information from other sources such as the PubChem Bioassay database [100, 101]. The latest version of the ChEMBL database counts over 1.8 million distinct compounds annotated with more than 15.2 million activity records on a total of more than 12,000 targets. In our recent analysis, we found that approximately 16% (40,000) of known natural products are contained in ChEMBL [10].

6 Resources for Structural Data on Natural Products

The Cambridge Structural Database (CSD) [102] provides a wealth of information on the three-dimensional structures of small-molecule organic and metal-organic compounds. Currently, the database is approaching the milestone of storing 1 million structures derived by X-ray and neutron diffraction analysis.

Structural information of natural products bound to their biomacromolecular targets is available from the Protein Data Bank (PDB) [103] but remains sparse. We found that for approximately 2000 natural products at least one X-ray crystal structure in complex with a biomacromolecule is deposited in the PDB [13]. A small number of structures of protein-bound macrocyclic natural products are also available [104].

7 Conclusions

During the last few years, the chemical, biological, and structural information available on natural products has increased substantially. Today, the molecular structures of several hundred thousand natural products are available from a large number of different sources. In particular, natural products from botanical sources are to a large part covered by subscription-free resources that permit bulk export or download of data, allowing an array of different cheminformatics methods to be

employed in the context of drug discovery. It is important to mention that the quality and quantity of the information provided by the individual sources vary substantially. For example, not all sources provide information on stereochemical properties, which in fact are often incomplete or inaccurate for natural products anyway. To the best of our knowledge, there have been no systematic studies on the quality of the data provided by natural product databases. This would, of course, be an important aspect to examine further.

Measured data on biological activities and ADME properties are becoming increasingly available, whereas structural information on natural products bound to their biomacromolecular target remain sparse. The bottleneck for drug discovery continues to be the availability of material for experimental testing. It is estimated that only about 10% (25,000) of all known natural products are readily obtainable from commercial and other sources. However, a substantially higher number of natural product-like compounds are readily obtainable.

In the coming years, we expect a further increased growth rate for chemical, biological, and structural data on natural products. In particular, we expect resources providing free access and bulk data download to play an ever more important role. One major challenge is to develop strategies for the sustainability of such valuable sources. What is seen today, unfortunately, is that many databases are no longer maintained after they have been reported in the scientific literature, and there are many examples of resources that go offline even within 1 year after their launch. This phenomenon is, of course, not specific to natural product databases but part of a general and largely unsolved problem.

Despite the remaining challenges, the large amount of data on natural products available today enables investigators to effectively employ computational methods and make substantial contributions to natural product-based drug discovery.

Funding Sources Ya Chen is supported by the China Scholarship Council (201606010345). Johannes Kirchmair is supported by the Bergen Research Foundation (BFS)—grant no. BFS2017TMT01.

References

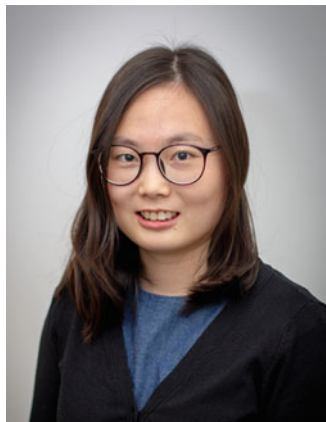
1. Newman DJ, Cragg GM (2016) Natural products as sources of new drugs from 1981 to 2014. *J Nat Prod* 79:629
2. Harvey AL, Edrada-Ebel R, Quinn RJ (2015) The re-emergence of natural products for drug discovery in the genomics era. *Nat Rev Drug Discov* 14:111
3. Stratton CF, Newman DJ, Tan DS (2015) Cheminformatic comparison of approved drugs from natural product versus synthetic origins. *Bioorg Med Chem Lett* 25:4802
4. Ertl P, Schuffenhauer A (2008) Cheminformatics analysis of natural products: lessons from Nature inspiring the design of new drugs. In: Petersen F, Amstutz R (eds) *Natural compounds as drugs*, vol II. Birkhäuser Verlag, Basel, p 217
5. Chen H, Engkvist O, Blomberg N, Li J (2012) A comparative analysis of the molecular topologies for drugs, clinical candidates, natural products, human metabolites and general bioactive compounds. *Med Chem Commun* 3:312

6. Feher M, Schmidt JM (2003) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J Chem Inf Comput Sci* 43:218
7. Clemons PA, Bodycombe NE, Carrinski HA, Wilson JA, Shamji AF, Wagner BK, Koehler AN, Schreiber SL (2010) Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc Natl Acad Sci U S A* 107:18787
8. Bisson J, McAlpine JB, Friesen JB, Chen S-N, Graham J, Pauli GF (2016) Can invalid bioactives undermine natural product-based drug discovery? *J Med Chem* 59:1671
9. Rodrigues T (2017) Harnessing the potential of natural products in drug discovery from a cheminformatics vantage point. *Org Biomol Chem* 15:9275
10. Chen Y, de Bruyn Kops C, Kirchmair J (2017) Data resources for the computer-guided discovery of bioactive natural products. *J Chem Inf Model* 57:2099
11. DNP – Dictionary of Natural Products (2019) <http://dnp.chemnetbase.com>
12. Reaxys; Elsevier: New York (2019) <https://www.reaxys.com>
13. Chen Y, Garcia de Lomana M, Friedrich N-O, Kirchmair J (2018) Characterization of the chemical space of known and readily obtainable natural products. *J Chem Inf Model* 58:1518
14. Laatsch H (2017) AntiBase: the natural compound identifier. Wiley-VCH, Weinheim
15. AntiBase (2019) <https://application.wiley-vch.de/stmdata/antibase.php>
16. Banerjee P, Erehman J, Gohlke B-O, Wilhelm T, Preissner R, Dunkel M (2014) Super natural II – a database of natural products. *Nucleic Acids Res* 43:D935
17. SuperNatural II (2019) http://bioinf-applied.charite.de/supernatural_new
18. Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X (2013) Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One* 8:e62839
19. Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, Wang Y, Tan Y, Gao D, Wang B, Chen Z, Chen W, Jiang YY, Chen YZ (2018) NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res* 46:D1217
20. NPASS – Natural Product Activity and Species Source Database (2019) <http://bidd2.nus.edu.sg/NPASS/index.php>
21. Zeng X, Zhang P, Wang Y, Qin C, Chen S, He W, Tao L, Tan Y, Gao D, Wang B, Chen Z, Chen W, Jiang YY, Chen YZ (2019) CMAUP: a database of collective molecular activities of useful plants. *Nucleic Acids Res* 47:D1118
22. CMAUP – Collective Molecular Activities of Useful Plants (2019) <http://bidd2.nus.edu.sg/CMAUP/index.html>. Accessed 17 Jan 2019
23. Natural Products Atlas (2019) <https://www.npatlas.org>
24. Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG (2017) Retrospective analysis of natural products provides insights for future discovery trends. *Proc Natl Acad Sci U S A* 114:5601
25. Hao M, Cheng T, Wang Y, Bryant SH (2013) Web search and data mining of natural products and their bioactivities in PubChem. *Sci China Chem* 56:1424
26. PubChem Substance (2019) <http://ncbi.nlm.nih.gov/pcsubstance>
27. Chen CY-C (2011) TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One* 6:e15939
28. TCM Database@Taiwan (2019) <http://tcm.cmu.edu.tw>
29. Huang L, Xie D, Yu Y, Liu H, Shi Y, Shi T, Wen C (2018) TCMID 2.0: a comprehensive resource for TCM. *Nucleic Acids Res* 46:D1117
30. Li B, Ma C, Zhao X, Hu Z, Du T, Xu X, Wang Z, Lin J (2018) YaTCM: Yet another Traditional Chinese Medicine Database for drug discovery. *Comput Struct Biotechnol J* 16:600
31. YaTCM – yet another traditional Chinese medicine database (2019) <http://cadd.pharmacy.nankai.edu.cn/yatcm/home>
32. Ehrman TM, Barlow DJ, Hylands PJ (2007) Phytochemical informatics of traditional Chinese medicine and therapeutic relevance. *J Chem Inf Model* 47:2316

33. Chem-TCM (2019) www.chemtcm.com
34. Kang H, Tang K, Liu Q, Sun Y, Huang Q, Zhu R, Gao J, Zhang D, Huang C, Cao Z (2013) HIM-herbal ingredients in-vivo metabolism database. *J Cheminform* 5:28
35. Ye H, Ye L, Kang H, Zhang D, Tao L, Tang K, Liu X, Zhu R, Liu Q, Chen YZ, Li Y, Cao Z (2011) HIT: linking herbal active ingredients to targets. *Nucleic Acids Res* 39:D1055
36. Mohanraj K, Karthikeyan BS, Vivek-Ananth RP, Chand RPB, Aparna SR, Mangalapandi P, Samal A (2018) IMPPAT: A curated database of Indian medicinal plants, phytochemistry and therapeutics. *Sci Rep* 8:4329
37. IMPPAT – Indian Medicinal Plants, Phytochem Therapeutics (2019) <https://cb.imsc.res.in/imppat>
38. DMNP – Dictionary of Marine Natural Products (2019) <http://dmnp.chemnetbase.com>
39. MarinLit (2019) <http://pubs.rsc.org/marinlit>
40. Lin Y-C, Wang C-C, Chen I-S, Jheng J-L, Li J-H, Tung C-W (2013) TIPdb: a database of anticancer, antiplatelet, and antituberculosis phytochemicals from indigenous plants in Taiwan. *Sci World J* 2013:736386
41. Tung C-W, Lin Y-C, Chang H-S, Wang C-C, Chen I-S, Jheng J-L, Li J-H (2014) TIPdb-3D: the three-dimensional structure database of phytochemicals from Taiwan indigenous plants. *Database* 2014:bau055
42. TIPdb – Taiwan Indigenous Plant Database (2019) <http://cwtung.kmu.edu.tw/tipdb>
43. Ntie-Kang F, Telukunta KK, Döring K, Simoben CV, A Moumbock AF, Malange YI, Njume LE, Yong JN, Sippl W, Günther S (2017) NANPDB: a resource for natural products from northern African sources. *J Nat Prod* 80:2067
44. NANPDB – Northern African Natural Products Database (2019) www.african-compounds.org/nanpdb
45. Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, Mbah JA, Mbaze LM, Sippl W, Efang SMN (2013) AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PLoS One* 8:e78085
46. Hatherley R, Brown DK, Musyoka TM, Penkler DL, Faya N, Lobb KA, Tasthan Bishop Ö (2015) SANCDB: a South African natural compound database. *J Cheminform* 7:29
47. SANCDB - South African Natural Compound Database (2019) <http://sancdb.rubi.ru.ac.za>
48. Ntie-Kang F, Nwodo JN, Ibezim A, Simoben CV, Karaman B, Ngwa VF, Sippl W, Adikwu MU, Mbaze LM (2014) Molecular modeling of potential anticancer agents from African medicinal plants. *J Chem Inf Model* 54:2433
49. Onguéné PA, Ntie-Kang F, Mbah JA, Lifongo LL, Ndom JC, Sippl W, Mbaze LM (2014) The potential of anti-malarial compounds derived from African medicinal plants. Part III: an in silico evaluation of drug metabolism and pharmacokinetics profiling. *Org Med Chem Lett* 4:6
50. Saldívar-González FI, Valli M, Andricopulo AD, da Silva Bolzani V, Medina-Franco JL (2018) Chemical space and diversity of the NuBBE database: a chemoinformatic characterization. *J Chem Inf Model* 59:74
51. Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2017) NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci Rep* 7:7215
52. Valli M, dos Santos RN, Figueira LD, Nakajima CH, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2013) Development of a natural products database from the biodiversity of Brazil. *J Nat Prod* 76:439
53. NuBBE – Núcleo de Bioensaios, Biossíntese e Ecofisiologia de Produtos Naturais (2019) <http://nubbe.iq.unesp.br/portal/nubbe-search.html>
54. Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL (2019) BIOFACQUIM: a Mexican compound database of natural products. *Biomolecules* 9:31
55. BIOFACQUIM (2019) <https://biofacquim.herokuapp.com>
56. Huang W, Brewer LK, Jones JW, Nguyen AT, Marcu A, Wishart DS, Oglesby-Sherrouse AG, Kane MA, Wilks A (2018) PAMDB: a comprehensive *Pseudomonas aeruginosa* metabolome database. *Nucleic Acids Res* 46:D575

57. PAMDB — *Pseudomonas aeruginosa* Metabolome Database (2019) <http://pseudomonas.umaryland.edu/PAMDB.htm>
58. Klementz D, Döring K, Lucas X, Telukunta KK, Erxleben A, Deubel D, Erber A, Santillana I, Thomas OS, Bechthold A, Günther S (2015) StreptomeDB 2.0 – an extended resource of natural products produced by Streptomycetes. *Nucleic Acids Res* 44:D509
59. Streptome DB (2019) www.pharmaceutical-bioinformatics.de/streptomedb
60. Choi H, Cho SY, Pak HJ, Kim Y, Choi J-Y, Lee YJ, Gong BH, Kang YS, Han T, Choi G, Cho Y, Lee S, Ryoo D, Park H (2017) NPCARE: database of natural products and fractional extracts for cancer regulation. *J Cheminform* 9:2
61. NPCARE – Database of Natural Products for Cancer Gene Regulation (2019) <http://silver.sejong.ac.kr/npcare>
62. Mangal M, Sagar P, Singh H, Raghava GPS, Agarwal SM (2013) NPACT: naturally occurring plant-based anti-cancer compound-activity-target database. *Nucleic Acids Res* 41:D1124
63. NPACT – Naturally Occurring Plant-based Anticancerous Compound-Activity-Target Database (2019) <http://crdd.osdd.net/raghava/npact>
64. Zhang R, Lin J, Zou Y, Zhang X-J, Xiao W-L (2018) Chemical space and biological target network of anti-inflammatory natural products. *J Chem Inf Model* 59:66
65. Yabuzaki J (2017) Carotenoids Database: structures, chemical fingerprints and distribution among organisms. Database 2017:bax004
66. Carotenoid Database (2019) <http://carotenoiddb.jp/>
67. Shen J, Xu X, Cheng F, Liu H, Luo X, Shen J, Chen K, Zhao W, Shen X, Jiang H (2003) Virtual screening on natural products for discovering active compounds and target information. *Curr Med Chem* 10:2327
68. Qiao X, Hou T, Zhang W, Guo S, Xu X (2002) A 3D structure database of components from Chinese traditional medicinal herbs. *J Chem Inf Comput Sci* 42:481
69. He M, Yan X, Zhou J, Xie G (2001) Traditional Chinese medicine database and application on the Web. *J Chem Inf Comput Sci* 41:273
70. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH (2016) PubChem substance and compound databases. *Nucleic Acids Res* 44:D1202
71. Sterling T, Irwin JJ (2015) ZINC 15 – ligand discovery for everyone. *J Chem Inf Model* 55:2324
72. ZINC15 (2019) <http://zinc15.docking.org>
73. Wang J, Zhou H, Han L, Chen X, Chen Y, Cao Z (2005) Traditional Chinese medicine information database. *Clin Pharmacol Ther* 78:92
74. Zhou J, Xie G, Yan X (2011) Encyclopedia of traditional Chinese medicines — molecular structures, pharmacological activities, natural sources and applications. Springer, Berlin
75. Novel Antibiotics Database (2019) <http://www.antibiotics.or.jp/journal/database/database-top.htm>
76. Nakamura Y, Afendi FM, Parvin AK, Ono N, Tanaka K, Hirai Morita A, Sato T, Sugiura T, Altaf-ul-Amin M, Kanaya S (2014) KNApSAcK metabolite activity database for retrieving the relationships between metabolites and biological activities. *Plant Cell Physiol* 55:e7
77. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-ul-Amin M, Darusman LK, Saito K, Kanaya S (2012) KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol* 53:e1
78. Chen Y, Stork C, Hirte S, Kirchmair J (2019) NP-Scout: Machine learning approach for the quantification and visualization of the natural product-likeness of small molecules. *Biomolecules* 9:43
79. Ambinter (2019) www.ambinter.com
80. GreenPharma (2019) www.greenpharma.com
81. AnalytiCon Discovery (2019) www.ac-discovery.com
82. Chengdu Biopurify Phytochemicals (2019) www.biopurify.com

83. Selleck Chemicals (2019) www.selleckchem.com
84. TargetMol (2019) www.targetmol.com
85. Medchem Express (2019) www.medchemexpress.com
86. InterBioScreen (2019) www.ibscreen.com
87. TimTec (2019) www.timtec.net
88. AK Scientific (2019) www.aksci.com
89. Natural Products Set IV of the Developmental Therapeutic Program (DTP), NCI/NIH (2019) http://dtp.cancer.gov/organization/dscb/obtaining/available_plates.htm
90. INDOFINE Chemical Company (2019) www.indofinechemical.com
91. Pharmeks (2019) www.pharmeks.com
92. Princeton BioMolecular Research (2019) www.princetonbio.com
93. MicroSource Discovery Systems (2019) www.msdiscovery.com
94. Specs (2019) www.specs.net
95. Molecular Operating Environment (MOE), version 2016.08; Chemical Computing Group ULC, Montreal, QC
96. Lucas X, Grüning BA, Bleher S, Günther S (2015) The purchasable chemical space: a detailed picture. *J Chem Inf Model* 55:915
97. Morgan HL (1965) The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. *J Chem Doc* 5:107
98. Marsault E, Peterson ML (2011) Macrocycles are great cycles: applications, opportunities, and challenges of synthetic macrocycles in drug discovery. *J Med Chem* 54:1961
99. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 42:D1083
100. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S, Shoemaker BA, Bolton E, Gindulyte A, Bryant SH (2012) PubChem's BioAssay Database. *Nucleic Acids Res* 40:D400–D412
101. Wang Y, Bryant SH, Cheng T, Wang J, Gindulyte A, Shoemaker BA, Thiessen PA, He S, Zhang J (2017) PubChem BioAssay: 2017 update. *Nucleic Acids Res* 45:D955
102. Groom CR, Bruno IJ, Lightfoot MP, Ward SC (2016) The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater* 72:171
103. Berman HM (2000) The protein data bank. *Nucleic Acids Res* 28:235
104. Friedrich N-O, Flachsenberg F, Meyder A, Sommer K, Kirchmair J, Rarey M (2019) Conformer: a novel method for the generation of conformer ensembles. *J Chem Inf Model*. <https://doi.org/10.1021/acs.jcim.8b00704>



Ya Chen is a Ph.D. student in the research group of Professor Johannes Kirchmair at the Center for Bioinformatics (ZBH) of the University of Hamburg. She received her bachelor's degree in pharmacy from Jilin University (2013) and her master's degree in medicinal chemistry from Peking University (2016). Her research is focused on the development and application of computational methods for the identification of bioactive natural products and the prediction of their biomacromolecular targets. Together with Christina de Bruyn Kops and Johannes Kirchmair, she has recently published analyses of the chemical space of natural products and the subset of natural products that are readily obtainable (*J Chem Inf Model*, 2017, 57:2099 and *J Chem Inf Model*, 2018, 58:1518). She has also developed machine-learning models that discriminate between natural products and synthetic molecules with high accuracy (*Biomolecules* 2019, 9:43).



Christina de Bruyn Kops is a Ph.D. student in the research group of Professor Johannes Kirchmair at the Center for Bioinformatics (ZBH) of the University of Hamburg. She holds bachelor's degrees in chemistry and mathematics from Rice University (2013) and a master's degree in bioinformatics from the University of Hamburg (2016). The focus of her research is on the development of computational methods to predict xenobiotic metabolism. She is also interested in the role of natural products in drug discovery.



Johannes Kirchmair is an Associate Professor in bioinformatics at the Department of Chemistry and the Computational Biology Unit (CBU) of the University of Bergen. He also is a Group Leader at the Center for Bioinformatics (ZBH) of the University of Hamburg. After earning his Ph.D. from the University of Innsbruck (2007), Dr. Kirchmair started his career as an Application Scientist at Inte:Ligand GmbH (Vienna) and as a University Assistant at his alma mater. In 2010, he joined BASF SE (Ludwigshafen) as a Postdoctoral Research Fellow. Thereafter, he worked as a Research Associate at the University of Cambridge (2010–2013) and ETH Zurich (2013–2014). From 2014 to 2018, Johannes held a Junior Professorship in applied bioinformatics at the University of Hamburg. Dr. Kirchmair has been a Visiting Professor or lecturer at the National Institute of Warangal (2016), the University of Cagliari (2017), and the University of Vienna (2018). His main research interests include the development and application of computational methods for the prediction of the biological activities, metabolic fate, and toxicity of small molecules in the context of drug discovery.

A Toolbox for the Identification of Modes of Action of Natural Products



Tiago Rodrigues

Contents

1	Introduction	73
2	Molecular Docking	75
2.1	Identification of Modes of Action with Docking	76
3	Pharmacophore Model-Based Screening	77
3.1	Identification of Modes of Action with Pharmacophore Models	79
4	Molecular Similarity Searches	81
4.1	Identification of Modes of Action Through Structural Similarity	83
5	Machine Learning Methods	83
5.1	Identification of Modes of Action Using Learning Algorithms	84
6	Outlook	91
	References	92

1 Introduction

Natural products have long played a leading role in successful chemical biology and drug discovery, providing chemotypes sufficiently tailored to serve as chemical probes, drug leads or, at the very least, as sources of inspiration for molecular design [1–4]. While the development of innovative chemistry has facilitated the access to new and more diverse natural products in amounts suitable for bioactivity screening [5], prioritizing target-based assays remains not only a bottleneck in drug discovery but is also troublesome [6]. In fact, screening natural products of interest in target-based assays is often motivated by a prior phenotype change observation induced by the studied natural product in cell-based assays, e.g., cancer cell growth inhibition [6, 7]. Typically, the effective development of such bioactive natural products as useful drug leads relies on the deconvolution of the phenotypic readout and correlation of the said phenotype with the engagement of any given drug target or targets

T. Rodrigues (✉)

Chemical Biology, Instituto de Medicina Molecular João Lobo Antunes, Lisbon, Portugal
e-mail: tiago.rodrigues@medicina.ulisboa.pt

© Springer Nature Switzerland AG 2019

A. D. Kinghorn, H. Falk, S. Gibbons, J. Kobayashi, Y. Asakawa, J.-K. Liu (eds.),

Progress in the Chemistry of Organic Natural Products, Vol. 110,

https://doi.org/10.1007/978-3-030-14632-0_3

73

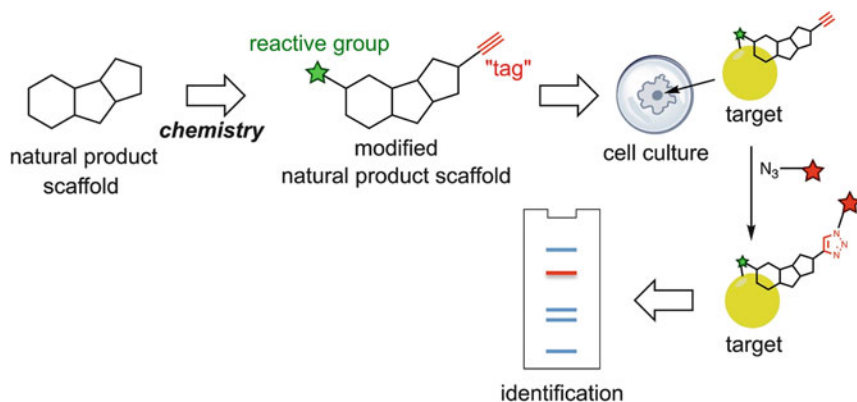


Fig. 1 Typical workflow for identifying drug targets through chemical proteomics approaches

[7]. It is now widely accepted that natural products, like small molecules, rarely are selective but engage dozens of related or unrelated targets [8], resulting in intricate pharmacology networks that might be explored in a drug discovery context [9, 10]. Crucially, such knowledge may bring benefits to the design of leads with lower probability of attrition and ultimately afford efficacious disease modulators.

Over the past few years, chemical proteomics (or chemoproteomics) has been established as the method of choice to identify binding counterparts for bioactive matter [7, 11]. In essence, the small molecule of interest is modified to incorporate a chemical handle prone to “tagging”; the modified chemical entity is then used to pull down proteins from cell lysates prior to subjecting such proteins to a downstream analytical method for identification (Fig. 1). In a recent prominent example, Cravatt and coworkers modified the diterpenoid ester ingenol mebutate—a first-in-class drug used for the treatment for actinic keratosis—to obtain a diazirine probe [12]. Using this photoreactive moiety the mitochondrial carnitine-acylcarnitine translocase SLC25A20 was identified as a functional target of ingenol mebutate. Despite the success in identifying a translocase as binding counterpart, membrane proteins are only seldomly identified as molecular targets using chemoproteomics as are proteins with low cellular expression [2]. Furthermore, the need for chemical modification of a molecule of interest may increase significantly chemical synthesis work, particularly in the case of natural products, and inadvertently disrupt the binding affinity towards relevant on- and off-targets [13]. Altogether, one may appreciate that the field of chemical proteomics is laborious, time consuming, and may require expensive equipment, and only provides motivated research hypotheses that must be validated with functional assays [6].

It is conceivable that in silico methods can provide viable alternatives to generate such motivated research hypotheses, yet within a fraction of time and resources spent. Virtual screening of an enumerated fraction of chemical space has been employed widely with vendor libraries as a means of accelerating hit discovery and prioritizing chemical matter for screening campaigns [14]. In contrast to

chemical proteomics, where target identification is a step downstream from phenotypic assays, *in silico* screening often focuses on a drug target for which ligands are sought [3]; only in the event of successful experimental validation of the predicted ligand-target relationship is the engagement of the target correlated with modulation of disease or adverse drug reactions [13, 15].

In this contribution, an overview will be provided and a discussion of strengths and limitations of computational methods that have been successfully employed for unveiling targets in the natural product realm. In particular, molecular docking and pharmacophore model-based strategies will be described as a means of accounting for three-dimensionality in scrutinizing potential drug targets for either natural products or synthetic small molecules. Importantly, with the advent of big data in biological and chemical sciences [16, 17], molecular docking and pharmacophore screening have become suboptimal approaches to process large volumes of information. In fact, the increasing computer power, storage capacity, and improved algorithms to analyze unstructured and sparse data, are setting the tone for a new era of cheminformatics where artificial intelligence promises to tackle some of the long-standing problems in molecular informatics and chemistry in general [17, 18]. As such, a special focus is given to emerging machine learning tools that leverage topological descriptors as a workhorse to building predictive models, and how such approaches can drive future chemical biology and early drug discovery programs. By comparing different tools, some of them accessible through webservers [8, 19], this contribution aims at being a reference work for the motivated selection of any given tool according to the goal of the project.

2 Molecular Docking

Molecular docking has become a standard means of screening virtual libraries within the realm of receptor-based methods [20, 21]. In short, such methods sample the ligand conformation space in a user-defined “box”/binding site in an attempt to predict the so-called “docking pose” and rationalize, on a molecular structure level, the activity that any compound might present against a given protein [20]. Thus, molecular docking software tools do not aim at identifying ready-made and optimized ligands, but rather discriminate relevant chemical features responsible for a molecular recognition event. These compounds and spatial arrangement of features might then be further tuned through medicinal chemistry to enhance binding affinity and, ideally, improve functional activity. Despite the simplicity of the concept and the existence of several user-friendly tools to carry out molecular docking studies, the researcher must bear in mind several caveats for proper data interpretation [20, 22, 23]. For instance, docking solely provides motivated research hypotheses or can rationalize them prior to experimental observations. Given that docking models account for only a snapshot of the protein in a conformational ensemble [21], they ought to be validated in biochemical studies (e.g., site mutagenesis) and the accuracy of the output is tightly connected to the quality of the protein X-ray

structure where docking is performed. Since X-ray structures represent electron density models, careful selection of the starting data is fundamental to avoid the exponential propagation of errors and inaccurate predicted poses. To this end, it is often advisable to select high-resolution structures (≤ 2.5 Å) and screen/correct amino acid residue rotamers, as assessed through Ramachandran plots [24–26].

While the search algorithms are generally able to find the correct pose [26], the scoring function that discerns the most likely and complementary ligand–target complex is often inaccurate at estimating the magnitude of the binding affinity. This is not a trivial task and endures as an active field of research. Binding affinity is best quantified by a free energy change between bound and unbound states as defined in Eq. (1):

$$\Delta G = \Delta H - T\Delta S \quad (1)$$

where G is the free energy of the ligand–receptor interaction, H is the enthalpy for the binding event, T is the absolute temperature, and S is the entropy. While the enthalpic contributions to binding can be both measured experimentally and modeled with some accuracy, this is not true for the entropic factor. Contributing to this is the limited information of protein flexibility in docking studies, the critical role of water molecules in mediating ligand–protein interactions or their displacement if unfavorable [27–29]. For example, inhibition of HIV protease by transition state mimetics occurs via displacement of a catalytic water molecule [30]. Usually, only a rough estimation of entropy change can be provided or else this is assumed to be identical in all cases. To mitigate this limitation, software tools such as WaterMap, can now evaluate statistically the position and the importance of each water molecule, and estimate if they are either structural or bulk solvent [31, 32]. Indeed, the physics-based modulation of water molecules can directly impact the entropic factor of binding for the ligand–target complex and provide more accurate modeling results [29]. Nonetheless, the scoring function data from mainstream molecular docking software tools should be analyzed with caution. These data serve well the purpose of generating a rank ordered list of ligands and help prioritize further investigations, but do not correlate with binding affinities.

2.1 Identification of Modes of Action with Docking

In keeping the drawbacks of molecular docking in mind, and analyzing the generated binding poses with healthy skepticism, it has been possible to deploy this technology with great effectiveness on natural products with the goal of unveiling putative binding partners that explain therapeutic effects and/or adverse drug reactions. For example, through inverse molecular docking, i.e., docking a single structure into several binding pockets of a large array of proteins, cyclooxygenase-2 (COX2; 56% inhibition at a concentration of $0.4 \mu\text{M}$) and peroxisome proliferator-activated

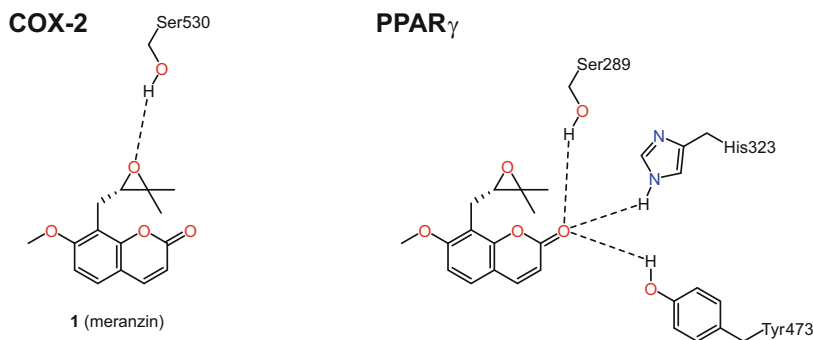


Fig. 2 Predicted interactions between the natural product meranzin (**1**) and cyclooxygenase-2 (COX2) and the peroxisome proliferator-activated receptor gamma (PPAR γ)

receptor gamma (PPAR γ ; active at concentrations above 10 μ M) were identified expeditiously as targets of meranzin (**1**) (Fig. 2). Importantly, this natural product displayed concentration-dependent effects and potencies comparable to indomethacin (COX2 ligand) and rosiglitazone (PPAR γ ligand) [33], suggesting that it could serve as a source of inspiration to design improved target effectors.

3 Pharmacophore Model-Based Screening

A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response. A pharmacophore does not represent a real molecule or a real association of functional groups, but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds toward their target structure. The pharmacophore can be considered the largest common denominator shared by a set of active molecules. This definition avoids a misuse often found in the medicinal chemistry literature, which consists of naming as pharmacophores simple chemical functionalities such as guanidines, sulfonamides, or dihydroimidazoles (formerly imidazolines), or typical structural skeletons such as flavones, phenothiazines, prostaglandins, or steroids. In summary, Wermuth's definition of 3D pharmacophores encompasses different regions of molecules in 3D space that encode steric and electronic properties and are responsible for molecular recognition. Through application of the molecular similarity principle, one may then assume that ligands with similar pharmacophore feature arrangements are likely to bind to the same targets. With this in mind, it is then possible to rapidly identify isofunctional molecules without explicitly comparing chemical structures, which may considerably speed up the search process, when compared to molecular docking. Moreover, pharmacophores are a convenient and physicochemically valid means of comparing molecules and perform scaffold

Table 1 Comparison of pharmacophore feature assignment schemes, by four popular software tools

Feature	LigandScout	MOE	Phase	Catalyst
H-bond	Acceptor and donor located on heavy atom	Acceptor and donor located on heavy atom	Donor located on hydrogen and acceptor on heavy atom	Acceptor and donor located on heavy atom
Lipophilic	Aromatic rings are recognized	Aromatic rings are not recognized	Aromatic rings are not recognized	Aromatic rings are recognized
Aromatic	Represented with plane orientation	Depends on pharmacophore scheme	Represented with plane orientation	Represented with plane orientation
Charge transfer	No explicit charges	Explicit charges	No explicit charges	No explicit charges

hopping, taking into account that entities with similar biological behavior can present disparate frameworks. As with the case of molecular docking, the output of pharmacophore model-based screening may vary considerably, depending on the software tool employed. Indeed, different tools present distinct pharmacophore feature assignment rules (Table 1), but all of them consider a tolerance zone that can be occupied by the atoms conferring a given property/feature.

It is good practice to take into account a range of different molecules binding to the same target being queried and generate several pharmacophore models for virtual screening purposes. Indeed, despite binding to the same target, it is not uncommon that modulators of a given target recognize different surface patches or recognize particular subpockets within a binding pocket. In this case, the ligands will modulate the same target through disparate binding modes. As such, it is sensible to cluster molecules in the reference ligand set by structural similarity, and generate as many models as the number of chemotypes, if there is no compelling evidence of identical modes of binding. Pharmacophore models may be computed either by performing multiple ligand alignment or ideally, by superimposing known bioactive conformations. In doing so, one is more likely to build relevant models for virtual screening. With such data in hand, features and their tolerance spheres can then be calculated automatically. As in the case of the reference ligand (training) set, conformers must be calculated and stored for the search (test) ligand set. While searching for matches to the pharmacophore model within a conformer set is not a computationally expensive task, the same cannot be said regarding the conformer generation routine. A force field must be selected, the potential energy of each ligand minimized and a user-defined array of energetically distinct conformers assembled. One may intuitively consider that an accurate three-dimensional representation of the ligands is key to the successful use of pharmacophore models. However, it has been suggested that the impact of the bioactive conformation on the overall database enrichment is limited [34–36]. Nevertheless, the computation of reasonable low-energy conformers is an important and a difficult task [37, 38]. This consideration is particularly true for natural products [39], for which the high content of stereogenic centers can

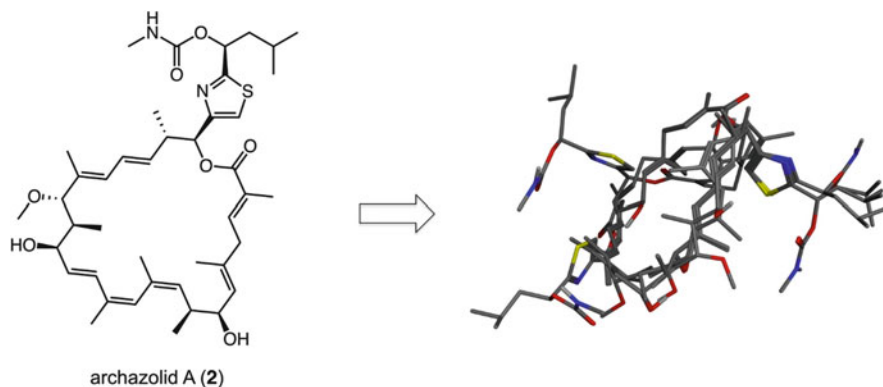


Fig. 3 Structure of a natural product and the superimposition of energy-minimized conformers, as computed with MOE (Chemical Computing Group, Canada). Data show that several distinct conformers are generated from the same structure

lead to several inaccurate and/or irrelevant conformers, as exemplified by archazolid A (2) (Fig. 3). Fortunately, as laborious as the conformer generation step may be, each search database needs to undergo the process only once—the output can be stored for future use. Taken together, and considering the caveats of conformer generation, pharmacophore model-based virtual screening is a viable alternative to molecular docking for rapid retrieval of hits.

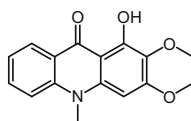
3.1 Identification of Modes of Action with Pharmacophore Models

Using 3D pharmacophore models, Rollinger and coworkers have successfully interrogated binding and engagement of targets by different natural products and their analogues. For example, a range of metabolites from common rue, *Ruta graveolens* (Plate 1), were screened against a panel of more than 2000 pharmacophore models to prioritize biochemical assays and experimentally confirm arborinine (3) (Fig. 4) and rutamarin (4) as inhibitors of the human rhinovirus coat protein and the G-protein coupled cannabinoid-2 receptor, respectively [40]. Moreover, hecogenin (5) isolated from the sisal plant, *Agave sisalana*, the labdane diterpenoid hispanolone (6), from *Ballota africana*, and lasalocid (7), from *Streptomyces lasaliensis*, have been identified as modulators of 11β -hydroxysteroid dehydrogenase [41], whereas several depside/depsidones, including perlatolic acid (8) from *Pertusaria globularis* and physodic acid (9) from *Pseudevernia furfuracea* were associated with inhibition of microsomal prostaglandin E2 synthase-1 [42]. Finally, PPAR γ was identified as target for biphenyl-based natural products, such as dieugenol (10) from aged clove basil (*Ocimum gratissimum*), magnolol (11) from the cortex of *Magnolia officinalis*,



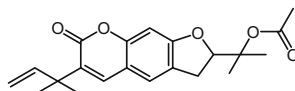
Plate 1 *Ruta graveolens*. Photograph: Jörg Hempel, Creative Commons 3.0

Human Rhinovirus Coat Protein



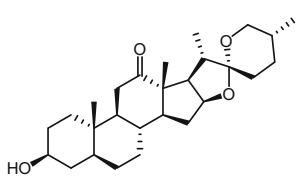
3 (arborinine)

Cannabinoid-2 receptor

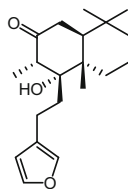


4 (rutamarin)

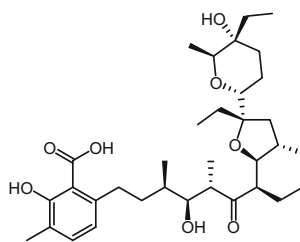
11 β -Hydroxysteroid Dehydrogenase (11 β -HSD)



5 (hecogenin)



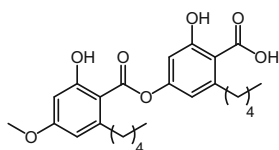
6 (hispanolone)



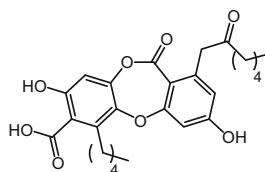
7 (lasalocid)

Fig. 4 Structures of natural products deorphanized through pharmacophore model-based virtual screening

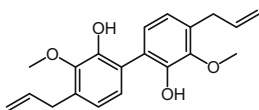
Microsomal Prostaglandin E2 Synthase-1 (mPGES-1)



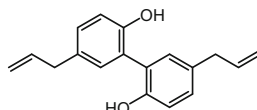
8 (perlatolic acid)



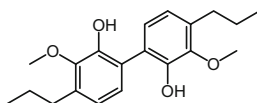
9 (physodic acid)

Peroxisome Proliferator-Activated Receptor Gamma (PPAR_γ)

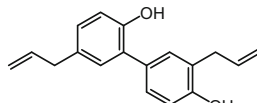
10 (dieugenol)



11 (magnolol)



12 (tetrahydrodieugenol)



13 (honokiol)

Fig. 4 (continued)

tetrahydrodieugenol (**12**) from the flowers of *Syzygium aromaticum*, and honokiol (**13**) also from the cortex of *M. officinalis* [43, 44].

4 Molecular Similarity Searches

Both molecular docking and 3D pharmacophore screening have been applied with great effectiveness to unveil putative binding counterparts for natural products. However, they rely on the computation of meaningful conformations; as discussed above, which is a particularly challenging endeavor. Additionally, these methods persist in being computationally expensive and arguably are of limited throughput.

In contrast to 3D methods, topological (2D) approaches offer viable alternatives of comparable accuracy, yet at a fraction of computational cost and speed [45]. Importantly, from a target inference point of view, the use of topological descriptors is well motivated, as similar ligands (and hence a similar resulting descriptor vector) are likely to bind to identical targets [19]. Thus, using appropriate descriptors/features to compare and correlate small molecules is key for success. Despite the high interest in designing efficient and highly informative descriptors, e.g., the extended 3-dimensional fingerprint (E3FP) that encodes stereochemical information [46], some methods remain mainstream. Among them is the use of physicochemical

descriptors, MACCS keys, or extended connectivity fingerprints (ECFPs) of different correlation diameters. Irrespective of the approach undertaken, the goal is to translate molecular structure into computable units that can be compared by one of several available metrics. Arguably, the Tanimoto-Jacquard coefficient/index [Eq. (2)] is the most widely employed metric to compare fingerprints, but others, such as dice similarity and Euclidean or Manhattan distances [Eq. (3)] have equally found applicability in cheminformatics to assess similarity between distinct molecules [47, 48]. The Tanimoto coefficient computes a value between zero and one to quantify the fingerprint similarity. A value of zero means complete dissimilarity between fingerprints of molecules under comparison, whereas a value of one indicates full identity. Therefore, the higher the value, the more similar the molecules will be according to the chosen fingerprint. Although there is no hard cutoff for similarity, it is generally accepted that a value equal or higher than 0.7–0.8 is obtained for similar ligands. Notably, the Tanimoto coefficient will vary significantly, depending on the chosen fingerprint and the number of bits (a certain substructural element) selected to store structural information. This will critically influence the accuracy of the approach and the molecules prioritized for experimental validation.

$$T = \frac{c}{a + b - c} \quad (2)$$

where T is the Tanimoto coefficient, a and b are the numbers of bits set for molecules A and B, under comparison, and c is the number of common bits in the fingerprints of molecules A and B.

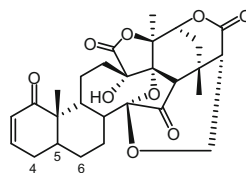
Euclidean and Manhattan distances can be computed through the Minkowski metric D , according to the formula (3):

$$D = \sqrt[p]{\sum_{i=1}^n |\mathbf{A}_i - \mathbf{B}_i|^p} \quad (3)$$

where n is the number of descriptor elements for molecules A and B. The formula affords the Manhattan and Euclidean distances for $p = 1$ or 2 , respectively.

In principle, any molecule with experimentally confirmed bioactivity against the target of interest can be used as starting point (reference) for similarity searches. However, taking into account that the goal of the method is to retrieve hits from a search database, a high-affinity ligand is a better motivated choice as reference molecule. Naturally, the selection of the descriptors employed and the metric used to assess similarity are cornerstones for the success of a screening campaign [49]. A wealth of screening techniques and software is available (some of which is implemented in open-source pipelining tools like KNIME), and their proper selection depends on the goal, suitability and availability, among others [50–52]. Irrespective of the screening strategy, similarity (or distance) values are calculated and stored in a database, which are then sorted in order of decreasing similarity (or increasing distance) to the query/reference molecules. The rank ordered list is

Fig. 5 Structures of antiplasmodial physalins



- 14** (physalin B): Δ^5
15 (physalin D): 5α -OH, 6β -OH
16 (physalin F): $5\beta,6\beta$ -epoxy
17 (physalin G): Δ^4 , 6α -OH

provided as output for human inspection, wherein the molecule with the smallest distance or higher similarity is called the “nearest neighbor.”

4.1 Identification of Modes of Action Through Structural Similarity

The Similarity Ensemble Approach (SEA, SeaChange Pharmaceuticals; webserver: <http://sea.bkslab.org>) [19, 53–55] leverages the similarity search concept discussed herein, coupled to probabilistic models to ascertain the relevance of its predictions. Notably, SEA allows prioritizing drug targets for screening with speed unrivaled by the abovementioned 3D methods. Having been developed primarily to identify on- and off-targets for synthetic small molecules, one can expect that high rates of false-positive predictions are obtained for natural products for which the frameworks diverge from those in the reference ligand database (ChEMBL [56]). While a thorough proof-of-concept is warranted, there is encouraging evidence that SEA can also perform efficiently, with natural products such as physalins B, D, F, and G (14–17) having been associated successfully with antiplasmodial activity (Fig. 5) [57].

5 Machine Learning Methods

Machine (statistical) learning is an (re-)emerging technology in chemical biology and drug discovery, with the potential to reshape how fundamental science is performed [18]. Its greatest value resides in leveraging an increasing amount of chemical and biological data to identify patterns and establish correlations that are otherwise intractable to human analyses [58]. Indeed, the recent progress made in all of computer storage, hardware, and algorithms provides a platform to foster investigations using machine learning as research tool. As in other modeling techniques, e.g., traditional quantitative structure-activity/property relationships, clearly defining

the research question is key to allow an appropriate strategy selection. Moreover, a certain amount of quality data is warranted to ensure that generalizable models are obtained for prospective deployment. To this end, tuning hyperparameters and performing cross-validation studies are equally important steps to assess whether the selected algorithm is under- or overfitting the training data. As a consequence of under- or overfit models, performance will be compromised when applied on related, yet previously unseen data. In brief, machine-learning technologies can be subdivided into three different categories, depending on the type of output and data requirements for learning:

1. Regression (supervised learning) if the output is a numeric value
2. Classification (supervised learning) if the output is a label
3. Clustering (unsupervised learning) if the algorithm associates data solely based on its structure

Independently of the method, all machine-learning approaches have proven useful in early drug discovery by streamlining processes and facilitating the design of relevant experiments. On one hand, regression and classification models have been employed prospectively for de novo design of small-molecule effectors [59], prediction of pharmacokinetics [60], prediction of drug-likeness [61], prediction of synthesis routes [62], optimization of chemical reactions [63], and conformational sampling [39], among many others. On the other hand, clustering methods have proven useful in the analysis of bioactivity landscapes [64, 65].

Given its utility for a number of tasks and the increase of bioactivity data for small molecules, machine learning has found applicability in research programs aiming at identifying targets for bioactive molecules of synthetic and natural origin [17]. Indeed, the need for minimal computational effort to afford statistically motivated research hypotheses renders machine learning as an attractive alternative to molecular docking and pharmacophore-based virtual screening.

5.1 Identification of Modes of Action Using Learning Algorithms

The Prediction of Activity Spectra for Substances (PASS) is available as an online tool (<http://www.pharmaexpert.ru/passonline/>) [66, 67], which uses topological fragment structure descriptors [68] and leverages a Bayesian-like method to infer > 2500 kinds of activities, including drug targets, for the queried molecules. Being a Bayes theorem-inspired method, PASS outputs probabilities of a studied molecule being active (P_a) or inactive (P_i). As such, a potentially interesting target for experimental validation will afford $P_a > P_i$, and the higher the difference, the more promising the target ought to be. To date, several marine sponge alkaloids have been scrutinized with PASS, and antitumor activity has been suggested for the great majority of them (80%) [69]. In addition to antitumor activity, PASS has also been

able to predict different kinds of activities for halitulins (**18**) from the sponge *Haliclona tulearensis* and betulin bishemiphthalate (**19**) a derivative of the triterpene betulin obtained from birch bark (Fig. 6). Thus, data suggest that these natural products may find broad applicability as therapeutics upon experimental confirmation of ligand–target correlations.

Considering the intricate frameworks in natural products and their dissimilarity to those entailed in synthetic molecules in reference datasets, one may argue that fingerprints and substructural descriptors are suboptimal to leverage confident target predictions in natural product space. Indeed, SEA and PASS were designed for synthetic entities, and may afford less accurate predictions than software tools tailored for natural products. To mitigate this limitation, the Chemically Advanced Template Search (CATS) computes topological pairwise correlations of atom types in a given molecule, up to a distance of 10 bonds [70, 71]. This simple pharmacophore descriptor provides a fuzzy and size-independent molecular representation, which has proven well suited for scaffold hopping and correlation of structurally dissimilar chemical entities. According to the CATS descriptors, feature pairs are expressed as the number of bonds along the shortest path connecting two non-hydrogen nodes in the molecular graph. Atoms are typed as one of six possible features: hydrogen bond donor, hydrogen bond acceptor, positively charged, negatively charged, lipophilic, and aromatic, resulting in a 210-dimensional vector (21 feature combinations \times 10 bonds) that can be employed to predict drug targets.

Taking advantage of the CATS descriptors, the Self-Organizing Maps (SOMs)-based prediction of drug equivalence (SPiDER) software [8, 72] uses a neural network heuristically inspired to achieve a weighted projection of the descriptor/chemical space onto a toroidal map in unsupervised fashion. To do so, the algorithm takes into account the structure of the input data and runs until convergence or for a user-defined number of epochs. SOMs, such as those implemented in SPiDER are of straightforward interpretation since the local neighborhoods in data are preserved in the projection, i.e., similar data points are located in the same or adjacent neurons. Besides the CATS descriptors, the SPiDER software also uses 2D physicochemical properties computed by MOE (Chemical Computing Group, Canada) to afford a complementary vantage point on data for both reference ligands and queries. Next,

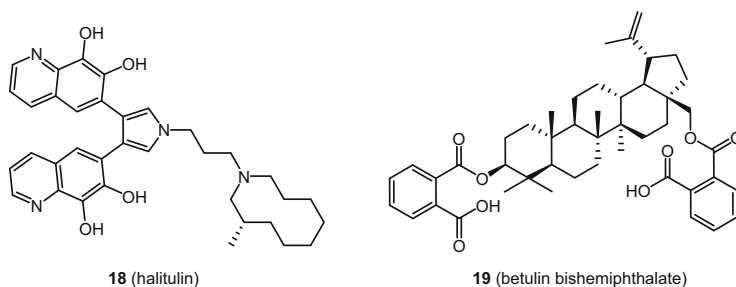


Fig. 6 Examples of a natural product **18** and a natural product derivative **19** studied with PASS

through arithmetical combination of the CATS and physicochemical descriptors-derived SOMs, and analyses of background distances between ligands, a consensus output is obtained together with a p -like value that allows assessment of the prediction significance (Fig. 7).

Although SPiDER was originally developed with the goal of inferring targets for de novo-designed small molecules [8], i.e., chemotypes displaying structural dissimilarity to their seed structures, Schneider and coworkers recognized that the same concept could be applied efficiently to deorphanize natural products and interrogate their polypharmacological profiles. In a first report, SPiDER was applied prospectively to the macrocyclic natural product archazolid A (**2**) [73]. As **2** differs considerably from ligands in the SPiDER reference database, only low confidence predictions could be obtained. The observation led to the deconvolution of the macrocyclic structure into its computationally generated fragments, assuming that the bioactivity fingerprint of **2** could be partly stored into those fragments and used subsequently as surrogate structures for SPiDER processing. Interestingly, different fragments afforded identical confident predictions, which were used to initiate biochemical assays. Compound **2** was confirmed as modulating COX2, PPAR γ , glucocorticoid receptor (GR), mPGES-1, and 5-lipoxygenase (5-LO), among others. Albeit not confirmed experimentally, modulation of these targets may contribute to its possible anticancer activity (Fig. 8). Similarly, the highly cytotoxic macrocycle dolicolide (**20**) from the Japanese sea hare (*Aplysia juliana*) was deorphanized as a nanomolar-potent prostanoid receptor 3 antagonist using synthetically motivated fragments to leverage a target prediction routine. Inhibition of the prostanoid receptor 3 may also be involved in cancer progression [74].

The SPiDER method has equally shown accuracy in identifying drug targets for fragment-like natural products. While (–)-sparteine (**21**) modulates the κ -opioid receptor ($EC_{50} = 245 \mu\text{M}$, Fig. 9) [3], isomacroid (**22**) was found to be an inhibitor of the platelet-derived growth factor receptor alpha kinase (PDGFR α) without selectivity for the beta isoform, but with negligible effects against a panel of diverse kinases [15]. Through substitution of the imidazole ring to the *N*-methylpyrrole counterpart, activity against PDGFR α was abrogated, which indicated the paramount role of the imidazole moiety as a hinge-binding motif. Indeed, compound **22** is a substructure of a single-digit nanomolar PDGFR β inhibitor developed by the pharmaceutical industry [75], further attesting to the validation of natural products as starting points for hit-to-lead optimization programs. In another case study, graveolinin (**23**) was identified as a COX2 and serotonin 5-HT $_{2B}$ modulator. Indeed,

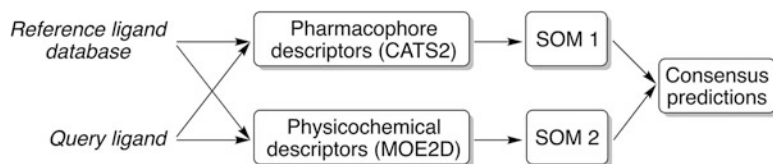


Fig. 7 Schematics of the SPiDER method workflow

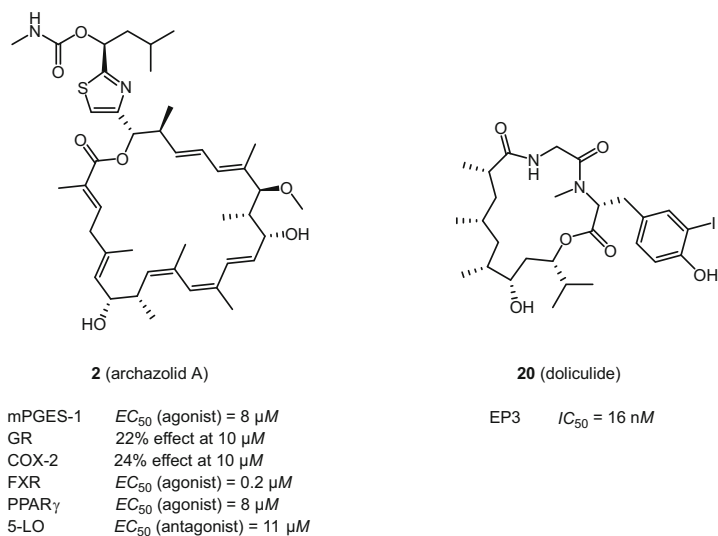


Fig. 8 Structures of archazolid A (**2**) and dolicolide (**20**) and bioactivities identified by SPiDER

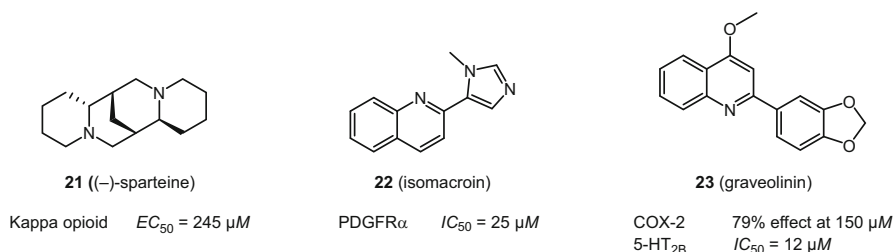


Fig. 9 Structures of fragment-like natural products deorphanized with SPiDER

inhibition of COX2 may explain the antiplatelet aggregation effect displayed by extracts of *Ruta graveolens* (Plate 1), for which the major constituent is **23** [76]. Importantly, despite the structural dissimilarity to typical COX2 inhibitors, a similar pharmacophore can explain the prediction made by SPiDER, and suggests that potent COX2 inhibitors inspired by **23** can be developed.

Finally, (-)-englerin A (**24**) (Fig. 10), a known renal antitumor cell agent from the African plant *Phyllanthus engleri*, which increases intracellular calcium concentration through activation of the transient receptor potential channel canonical 4 and 5 (TRPC4/5) [77, 78] was suggested as a voltage-gated calcium Ca $_v$ 1.2 channel ligand [79]. As has occurred for **2** and **20**, prediction of targets with the full natural product structure afforded only few confident predictions. To augment the number of confidently predicted targets, the authors used piperlongumine (**25**) from *Piper longum* as a pharmacophore surrogate for SPiDER, assuming that targets inferred

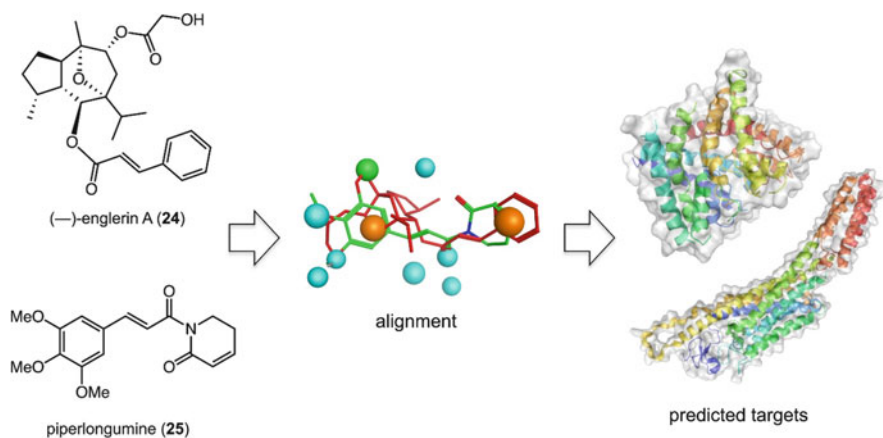


Fig. 10 (–)-Englerin A (**24**) and piperlongumine (**25**) display pharmacophore feature commonalities that allow cross-structure target inference. Cyan = hydrogen bond donor/acceptor; green = lipophilic; orange = aromatic and/or sp^2 hybridized

for alkaloid **25** would equally represent motivated research hypotheses for **24**, from a cheminformatics vantage point. A range of biochemical and cell-based assays confirmed that **24** moderately antagonized $Ca_v1.2$ channels ($IC_{50} = 6 \mu M$). Despite the low relevance of the finding to further explain the antitumor activity of **24**, the study afforded a rationale to graft natural product-derived fragments and tailor $Ca_v1.2$ modulators. Interestingly, SPiDER was also able to predict TRP channels as binding counterparts for **24**, which could significantly speed up target exploration studies. The result from the pseudo-prospective evaluation of **24** is in line with the observation that natural products are privileged ligands of TRP channels [80]. Altogether, the validation of **24** as a calcium channel modulator provides another example of the utility of machine learning in identifying membrane proteins as targets of bioactive matter.

The SPiDER method has recently been replicated to afford the Target Inference Generator (TIGER) tool that also leverages a consensus of two SOMs, but slightly modified CATS descriptors, i.e., without charged features, and a disparate statistical approach. By encoding ligand–target relationships, TIGER is capable of performing qualitative predictions of up to 331 targets [81], among which orexin 1/2, glucocorticoid and cholecystokinin-2 receptors that were experimentally validated for the marine natural product (\pm)-marinopyrrole A (**26**) isolated from a *Streptomyces* sp. (Fig. 11). In an additional prospective application of TIGER, resveratrol (**27**) was predicted and experimentally confirmed to modulate the estrogen receptor β ($ER\beta$, $K_i = 0.4 \mu M$) with a reasonable degree of selectivity over its α counterpart ($ER\alpha$, $K_i = 4 \mu M$) [82].

Built with the goal of scrutinizing the qualitative SPiDER predictions, Rodrigues et al. reported the Drug–Target Relationship Predictor (DEcRyPT) software tool [13] that uses random forest technology to predict affinity values for targets of interest. In

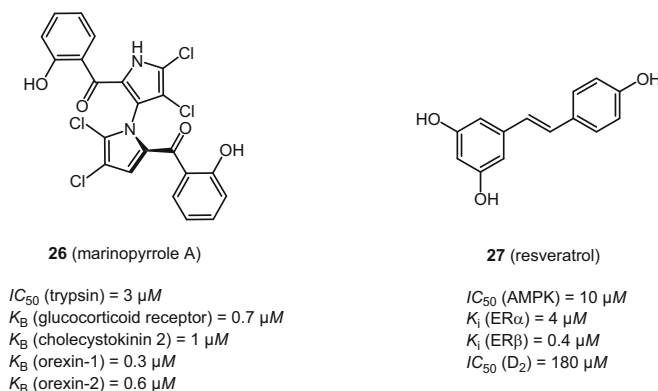


Fig. 11 Structures of natural products that have been studied by the TIGER method

short, random forest models leverage the individual predictions of a user-defined number of decision trees, built with only a subset of all data. As such, each decision tree functions as a weak estimator, but as an ensemble, more robust and reliable predictions are made, with the added value of theoretically reducing over-fitting and improving the model generalizability. The tool DEcRyPT was built with curated and transformed bioactivity data as collected from ChEMBL, v.22 [56] and the CATS2 descriptors [70]. Applying DEcRyPT to β -lapachone (**28**) (Fig. 12), originally isolated from the heartwood of the South American Lapacho tree (*Tabebuia avellanedae*), 5-LO emerged as potential target of interest. Using multiple cell-free assays, the authors confirmed that **28** must be converted to its hydroquinone form, which acts as a nanomolar inhibitor of 5-LO ($IC_{50} = 240$ nM). Importantly, the authors ruled out unspecific inhibition by colloidal aggregates [83], by confirming inhibitory activity of 5-LO independently of the presence or absence of Triton X-100. Moreover, inhibition of 5-LO was comparable in cell-free and whole-cell assays, confirming the absence of permeability issues that could hamper further exploratory work on this chemotype. Unexpectedly, compound **28** displayed selectivity for 5-LO over its congeners 15- and 12-LO, which suggests binding to an allosteric site (Fig. 12). Elaboration of these preliminary findings showed that in fact, compound **28** does not compete with the natural 5-LO ligand—arachidonic acid—nor is a general metal chelator. Conversely, the potency of **28** is reduced significantly in competition assays with phosphatidylcholine, which binds at the interface of the catalytic and C2-like domains. Finally, inhibition of 5-LO by the hydroquinone form of **28** (Fig. 12) could be correlated with the antitumor effects, as cells overexpressing 5-LO are more sensitive to the natural product.

In another application of DEcRyPT, secondary pharmacology was unveiled for DMP-1 (**29**)—a synthetic analogue of militarinone A (**30**) (Fig. 13) isolated from the mycelium of the entomogenous fungus *Paecilomyces militaris* [84]. The tool DEcRyPT predicted potent modulation of the cannabinoid receptor 1 (CB1) by **29** (predicted affinity of 0.16 μM) with high confidence, which was confirmed experimentally by determining functional antagonism with a potency of 0.32 μM , and

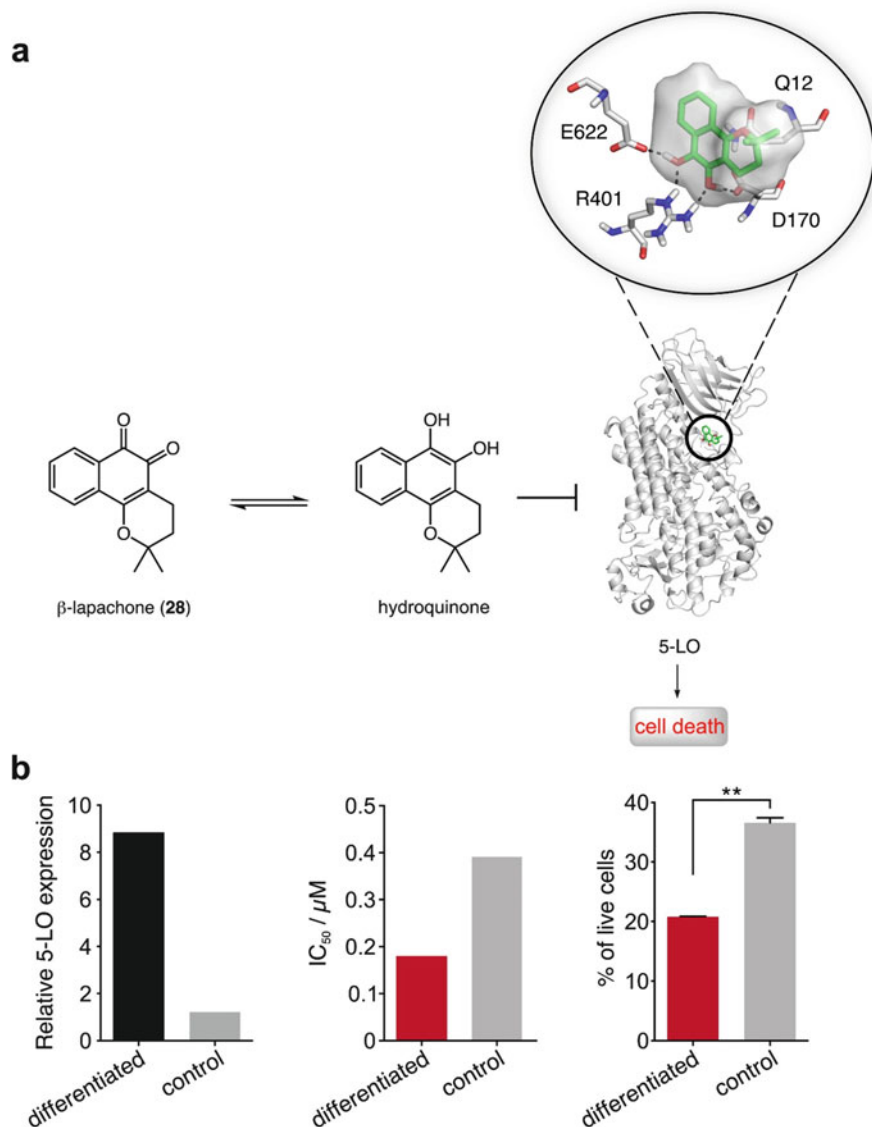


Fig. 12 Mechanism of anticancer activity of β -lapachone (**28**). (a) Natural product **28** is converted in the intracellular compartment to the corresponding hydroquinone, which is a potent, reversible, allosteric inhibitor of 5-lipoxygenase (5-LO). (b) Differentiated HL-60 cell line overexpresses 5-LO (left) and are more sensitive to **28** (middle and right). IC_{50} (differentiated) = $0.18 \mu M$; IC_{50} (control) = $0.39 \mu M$. Percentage of live HL-60 cells in the differentiated and control groups when treated with $0.5 \mu M$ of **28**. $**p < 0.005$ (two-tailed *t*-Student test)

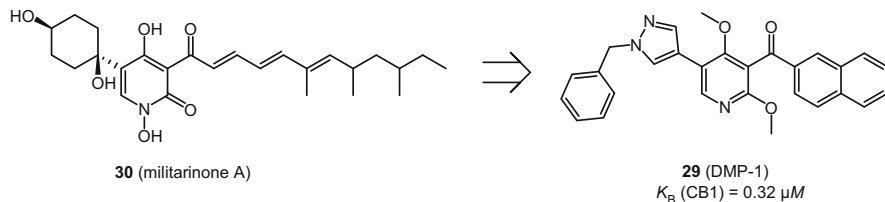


Fig. 13 Structure of militarinone A (**30**) and its derivative DMP-1 (**29**)

displacement of a radiolabeled ligand with a K_i value of 3.2 μ M [84]. Again, the identification of a trans-membrane protein was facilitated by machine intelligence, which could have hardly been done through chemical proteomics.

6 Outlook

Target identification and deconvolution of phenotypic readouts is an important step in early discovery programs. While this is a challenging task for synthetic small molecules, the difficulty is typically magnified for natural products, given the poorer synthetic accessibility and troublesome derivatization tendencies. However, with such knowledge in hand, developing bioactive natural products and designing analogues may be facilitated and assisted by state-of-the-art computational technologies. In this contribution, different *in silico* methods that can be of utility to unveil pharmacology of natural products have been discussed, and in a broader sense any small molecule of interest, by generating motivated research hypotheses for confirmation in biochemistry laboratories.

There is no universal best method and both 3D and 2D approaches can be deployed efficiently by keeping in mind their caveats and limitations. Still, any computational method is certain to fail occasionally even when properly employed, but more often when applied outside its domain of applicability. However, there is compelling evidence that the accuracy and scope of computational methods are improving considerably. This offers great prospects for more successful case studies in the deconvolution of modes of action and biochemical liabilities of natural products. Much of the current enthusiasm is spearheaded by the emergence of big data, faster computers, and more efficient algorithms for pattern recognition, which parallels the need for sustainable drug discovery. Machine learning is primed to analyze large volumes of data; these algorithms will equally benefit from high-quality negative data, which historically tends to be neglected. With the rise of digital chemistry, it is expected that laborious tasks such as target identification will be increasingly automated, thus opening new avenues for probabilistic drug discovery.

Acknowledgments Tiago Rodrigues is a Marie Skłodowska-Curie Fellow (Grant 743640) and acknowledges the FCT/FEDER (02/SAICT/2017, Grant 28333) for funding.

References

1. van Hattum H, Waldmann H (2014) Biology-oriented synthesis: harnessing the power of evolution. *J Am Chem Soc* 136:11853
2. Rodrigues T (2017) Harnessing the potential of natural products in drug discovery from a cheminformatics vantage point. *Org Biomol Chem* 15:9275
3. Rodrigues T, Reker D, Schneider P, Schneider G (2016) Counting on natural products for drug design. *Nat Chem* 8:531
4. Wetzel S, Bon RS, Kumar K, Waldmann H (2011) Biology-oriented synthesis. *Angew Chem Int Ed* 50:10800
5. Baran PS (2018) Natural product total synthesis: as exciting as ever and here to stay. *J Am Chem Soc* 140:4751
6. Laraia L, Waldmann H (2017) Natural product inspired compound collections: evolutionary principle, chemical synthesis, phenotypic screening, and target identification. *Drug Discov Today Technol* 23:75
7. Laraia L, Robke L, Waldmann H (2018) Bioactive compound collections: from design to target identification. *Chem* 4:705
8. Reker D, Rodrigues T, Schneider P, Schneider G (2014) Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc Natl Acad Sci U S A* 111:4067
9. Hopkins AL (2007) Network pharmacology. *Nat Biotechnol* 25:1110
10. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* 4:682
11. Moellering RE, Cravatt BF (2012) How chemoproteomics can enable drug discovery and development. *Chem Biol* 19:11
12. Parker CG, Kuttruff CA, Galmozzi A, Jørgensen L, Yeh CH, Hermanson DJ, Wang Y, Artola M, McKerrall SJ, Joslyn CM, Nørremark B, Dünstl G, Felding J, Saez E, Baran PS, Cravatt BF (2017) Chemical proteomics identifies SLC25A20 as a functional target of the ingenol class of actinic keratosis drugs. *ACS Cent Sci* 3:1276
13. Rodrigues T, Werner M, Roth J, da Cruz EHG, Marques MC, Akkapeddi P, Lobo SA, Koeberle A, Corzana F, da Silva Júnior EN, Werz O, Bernardes GJL (2018) Machine intelligence decrypts β -lapachone as an allosteric 5-lipoxygenase inhibitor. *Chem Sci* 9:6899
14. Schneider G (2010) Virtual screening: an endless staircase? *Nat Rev Drug Discov* 9:273
15. Rodrigues T, Reker D, Kunze J, Schneider P, Schneider G (2015) Revealing the macromolecular targets of fragment-like natural products. *Angew Chem Int Ed* 54:10516
16. Singh G, Schulthess D, Hughes N, Vannieuwenhuysse B, Kalra D (2018) Real world big data for clinical research and drug development. *Drug Discov Today* 23:652
17. Mayr A, Klambauer G, Untertiner T, Steijaert M, Wegner JK, Ceulemans H, Clevert D-A, Hochreiter S (2014) Biology-oriented synthesis: harnessing the power of evolution. *J Am Chem Soc* 136:11853
18. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today* 20:318
19. Keiser MJ, Roth BL, Armbruster BN, Emsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 25:197
20. Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct* 32:335

21. Totrov M, Abagyan R (2008) Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol* 18:178
22. Pagadala NS, Syed K, Tuszynski J (2017) Software for molecular docking: a review. *Biophys Rev* 9:91
23. Meng XY, Zhang HX, Mezei M, Cui M (2011) Molecular docking: a powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des* 7:146
24. Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys J* 101:2525
25. Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66:12
26. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, Mortenson PN, Murray CW (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* 50:726
27. Roberts BC, Mancera RL (2008) Ligand-protein docking with water molecules. *J Chem Inf Model* 48:397
28. Filippakopoulos P, Qi J, Picaud S, Shen Y, Smith WB, Fedorov O, Morse EM, Keates T, Hickman TT, Felletar I, Philpott M, Munro S, McKeown MR, Wang Y, Christie AL, West N, Cameron MJ, Schwartz B, Heightman TD, La Thangue N, French CA, Wiest O, Kung AL, Knapp S, Bradner JE (2010) Selective inhibition of BET bromodomains. *Nature* 468:1067
29. Ladbury JE (1996) Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem Biol* 3:973
30. Brik A, Wong CH (2003) HIV-1 protease: mechanism and drug discovery. *Org Biomol Chem* 1:5
31. Bertoldo JB, Rodrigues T, Dunsmore L, Aprile FA, Marques MC, Rosado LA, Boutoureira O, Steinbrecher TB, Sherman W, Corzana F, Terenzi H, Bernardes GJL (2017) A water-bridged cysteine-cysteine redox regulation mechanism in bacterial protein tyrosine phosphatases. *Chem* 3:665
32. Cappel D, Sherman W, Beuming T (2017) Calculating water thermodynamics in the binding site of proteins – applications of WaterMap to drug discovery. *Curr Top Med Chem* 17:2586
33. Do Q, Lamy C, Renimel I, Sauvan N, André P, Himbert F, Morin-Allory L, Bernard P (2007) Reverse pharmacognosy: identifying biological properties for plants by means of their molecule constituents: application to meranzin. *Planta Med* 73:1235
34. Renner S, Schwab CH, Gasteiger J, Schneider G (2006) Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors. *J Chem Inf Model* 46:2324
35. Zhang Q, Muegge I (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J Med Chem* 49:1536
36. Hawkins PC, Skillman AG, Nicholls A (2007) Comparison of shape-matching and docking as virtual screening tools. *J Med Chem* 50:74
37. Gasteiger J, Rudolph C, Sadowski J (1990) Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput Method* 3:537
38. Chen JJ, Foloppe N (2008) Conformational sampling of druglike molecules with MOE and catalyst: implications for pharmacophore modeling and virtual screening. *J Chem Inf Model* 48:1773
39. Rupp M, Bauer MR, Wilcken R, Lange A, Reutlinger M, Boeckler FM, Schneider G (2014) Machine learning estimates of natural product conformational energies. *PLoS Comput Biol* 10: e1003400
40. Rollinger JM, JM Schuster D, Danzl B, Schwaiger S, Markt P, Schmidtke M, Gertsch J, Raduner S, Wolber G, Langer T, Stuppner H (2009) In silico target fishing for rationalized ligand discovery exemplified on constituents of *Ruta graveolens*. *Planta Med* 75:195
41. Vuorinen A, Nashev LG, Odermatt A, Rollinger JM, Schuster D (2014) Pharmacophore model refinement for 11 β -hydroxysteroid dehydrogenase inhibitors: search for modulators of intracellular glucocorticoid concentrations. *Mol Inf* 33:15

42. Bauer J, Waltenberger B, Noha SM, Schuster D, Rollinger JM, Boustie J, Chollet M, Stuppner H, Werz O (2012) Discovery of depsides and depsidones from lichen as potent inhibitors of microsomal prostaglandin E2 synthase-1 using pharmacophore models. *ChemMedChem* 7:2077
43. Fakhrudin N, Ladurner A, Atanasov AG, Heiss EH, Baumgartner L, Markt P, Schuster D, Ellmerer EP, Wolber G, Rollinger JM, Stuppner H, Dirsch VM (2010) Computer-aided discovery, validation, and mechanistic characterization of novel neolignan activators of peroxisome proliferator-activated receptor gamma. *Mol Pharmacol* 77:559
44. Atanasov AG, Wang JN, Gu SP, Bu J, Kramer MP, Baumgartner L, Fakhrudin N, Ladurner A, Malainer C, Vuorinen A, Noha SM, Schwaiger S, Rollinger JM, Schuster D, Stuppner H, Dirsch VM, Heiss EH (2013) Honokiol: a non-adipogenic PPARgamma agonist from Nature. *Biochim Biophys Acta* 1830:4813
45. Brown RD, Martin YC (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J Chem Inf Comput Sci* 37:1
46. Axen SD, Huang XP, Caceres EL, Gendele L, Roth BL, Keiser MJ (2017) A simple representation of three-dimensional molecular structure. *J Med Chem* 60:7393
47. Bajusz D, Racz A, Heberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminf* 7:20
48. Sheridan RP, Kearsley SK (2002) Why do we need so many chemical similarity search methods? *Drug Discov Today* 7:903
49. Bender A, Jenkins JL, Scheiber J, Sukuru SC, Glick M, Davies JW (2009) How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J Chem Inf Model* 49:108
50. Xue L, Bajorath J (2000) Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb Chem High Throughput Screen* 3:363
51. Bajorath J (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J Chem Inf Comput Sci* 41:233
52. Livingstone DJ (2000) The characterization of chemical structures using molecular properties. A survey. *J Chem Inf Comput Sci* 40:195
53. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijter MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KL, Edwards DD, Shoichet BK, Roth BL (2009) Predicting new molecular targets for known drugs. *Nature* 462:175
54. Loukine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Côté S, Shoichet BK, Urban L (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486:361
55. Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK (2008) Quantifying the relationships among drug classes. *J Chem Inf Model* 48:755
56. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR (2017) The ChEMBL database in 2017. *Nucleic Acids Res* 45:D945
57. Sá MS, de Menezes MN, Krettli AU, Ribeiro IM, Tomassini TC, Ribeiro dos Santos R, de Azevedo WF Jr, Soares MB (2011) Antimalarial activity of physalins B, D, F, and G. *J Nat Prod* 74:2269
58. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A (2018) Machine learning for molecular and materials science. *Nature* 559:547
59. Merk D, Friedrich L, Grisoni F, Schneider G (2018) De novo design of bioactive small molecules by artificial intelligence. *Mol Inf* 37:1700153
60. Kirchmair J, Williamson MJ, Afzal AM, Tyzack JD, Choy AP, Howlett A, Rydberg P, Glen RC (2013) FAsT MEtabolizer (FAME): a rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes. *J Chem Inf Model* 53:2896

61. Tian S, Wang J, Li Y, Xu X, Hou T (2012) Drug-likeness analysis of traditional Chinese medicines: prediction of drug-likeness using machine learning approaches. *Mol Pharmaceutics* 9:2875
62. Segler MHS, Preuss M, Waller MP (2018) Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555:604
63. Reker D, Bernardes GJL, Rodrigues T (2018) Evolving and nano data enabled machine intelligence for chemical reaction optimization. ChemRxiv: <https://doi.org/10.26434/chemrxiv.7291205.v7291201>
64. Reutlinger M, Rodrigues T, Schneider P, Schneider G (2014) Multi-objective molecular de novo design by adaptive fragment prioritization. *Angew Chem Int Ed* 53:4244
65. Rodrigues T, Hauser N, Reker D, Reutlinger M, Wunderlin T, Hamon J, Koch G, Schneider G (2015) Multidimensional de novo design reveals 5-HT_{2B} receptor-selective ligands. *Angew Chem Int Ed* 54:1551
66. Lagunin A, Stepanchikova A, Filimonov D, Poroikov V (2000) PASS: prediction of activity spectra for biologically active substances. *Bioinformatics* 16:747
67. Poroikov V, Filimonov D, Lagunin A, Glorizova T, Zakharov A (2007) PASS: identification of probable targets and mechanisms of toxicity. *SAR QSAR Environ Res* 18:101
68. Filimonov D, Poroikov V, Borodina Y, Glorizova T (1999) Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. *J Chem Inf Comput Sci* 39:666
69. Lagunin A, Filimonov D, Poroikov V (2010) Multi-targeted natural products evaluation based on biological activity prediction with PASS. *Curr Pharm Des* 16:1703
70. Reutlinger M, Koch CP, Reker D, Todoroff N, Schneider P, Rodrigues T, Schneider G (2013) Chemically Advanced Template Search (CATS) for scaffold-hopping and prospective target prediction for “orphan” molecules. *Mol Inf* 32:133
71. Schneider G, Neidhart W, Giller T, Schmid G (1999) “Scaffold-hopping” by topological pharmacophore search: a contribution to virtual screening. *Angew Chem Int Ed* 38:2894
72. Schneider P, Schneider G (2003) Collection of bioactive reference compounds for focused library design. *QSAR Comb Sci* 22:713
73. Reker D, Perna AM, Rodrigues T, Schneider P, Reutlinger M, Mönch B, Koeberle A, Lamers C, Gabler M, Steinmetz H, Müller R, Schubert-Zsilavec M, Werz O, Schneider G (2014) Revealing the macromolecular targets of complex natural products. *Nat Chem* 6:1072
74. Schneider G, Reker D, Chen T, Hauenstein K, Schneider P, Altmann KH (2016) Deorphanizing the macromolecular targets of the natural anticancer compound dolicolide. *Angew Chem Int Ed* 55:12408
75. Hicken EJ, Marmsater FP, Munson MC, Schlachter ST, Robinson JE, Allen S, Burgess LE, DeLisle RK, Rizzi JP, Topalov GT, Zhao Q, Hicks JM, Kallan NC, Tarlton E, Allen A, Callejo M, Cox A, Rana S, Klopfenstein N, Woessner R, Lyssikatos JP (2014) Discovery of a novel class of imidazo[1,2-*a*]pyridines with potent PDGFR activity and oral bioavailability. *ACS Med Chem Lett* 5:78
76. Wu TS, Shi L-S, Wang J-J, Iou S-C, Chang H-C, Chen Y-P, Kuo Y-H, Chang Y-L, Tenge C-M (2013) Cytotoxic and antiplatelet aggregation principles of *Ruta graveolens*. *J Chin Chem Soc* 50:171
77. Akbulut Y, Gaunt HJ, Muraki K, Ludlow MJ, Amer MS, Bruns A, Vasudev NS, Radtke L, Willot M, Hahn S, Seitz T, Ziegler S, Christmann M, Beech DJ, Waldmann H (2015) Englerin A is a potent and selective activator of TRPC4 and TRPC5 calcium channels. *Angew Chem Int Ed* 54:3787
78. Carson C, Raman P, Tullai J, Xu L, Henault M, Thomas E, Yeola S, Lao J, McPate M, Verkuyll JM, Marsh G, Sarber J, Amaral A, Bailey S, Lubicka D, Pham H, Miranda N, Ding J, Tang HM, Ju H, Tranter P, Ji N, Krastel P, Jain RK, Schumacher AM, Loureiro JJ, George E, Berellini G, Ross NT, Bushell SM, Erdemli G, Solomon JM (2015) Englerin A agonizes the TRPC4/C5 cation channels to inhibit tumor cell line proliferation. *PLoS One* 10:e0127498

79. Rodrigues T, Sieglitz F, Somovilla VJ, Cal PM, Galione A, Corzana F, Bernardes GJ (2016) Unveiling (-)-englerin A as a modulator of L-type calcium channels. *Angew Chem Int Ed* 55:11077
80. Rodrigues T, Sieglitz F, Bernardes GJ (2016) Natural product modulators of transient receptor potential (TRP) channels as potential anti-cancer agents. *Chem Soc Rev* 45:6130
81. Schneider P, Schneider G (2017) De-orphaning the marine natural product (\pm)-marinopyrrole A by computational target prediction and biochemical validation. *Chem Commun* 53:2272
82. Schneider P, Schneider G (2017) A computational method for unveiling the target promiscuity of pharmacologically active compounds. *Angew Chem Int Ed* 56:11520
83. Reker D, Bernardes GJL, Rodrigues T (2019) Computational advances in combating colloidal aggregation in drug discovery. *Nat Chem* 11(5):402–418
84. Robke L, Rodrigues T, Schröder P, Foley DJ, Bernardes GJL, Laraia L, Waldmann H (2018) Discovery of 2,4-dimethoxypyridines as novel autophagy inhibitors. *Tetrahedron* 74:4531



Tiago Rodrigues received an M.Sc. in pharmaceutical sciences (2006) and a Ph.D. in medicinal chemistry (2010) from the University of Lisbon, Portugal. He was a postdoctoral fellow at ETH Zürich (2011–2015) and then joined the Instituto de Medicina Molecular, Portugal, where he is currently a Staff Scientist. Since 2015, he is visiting Assistant Professor at the Faculty of Pharmacy, University of Lisbon. His research interests span a range of different disciplines, such as the development of drug delivery constructs, flow-assisted syntheses, and cheminformatics/machine learning for the deorphanization of natural products.

The Pharmacophore Concept and Its Applications in Computer-Aided Drug Design



Thomas Seidel, Doris A. Schuetz, Arthur Garon, and Thierry Langer

Contents

1	Introduction	99
2	The Pharmacophore Concept	100
2.1	Historical Background	100
2.2	Three-Dimensional Pharmacophores	103
2.2.1	Basic Interactions and Their Representation	105
2.2.2	Pharmacophore Elucidation	110
2.3	Application of Pharmacophores in Drug Design	114
2.3.1	Virtual Screening	114
2.3.2	Pharmacophore-Based De Novo Design	120
2.4	Current Research and Developments	123
2.4.1	Dynamic Pharmacophore Modeling and Virtual Screening	123
2.4.2	Pharmacophore-Based Interaction Fields	125
3	Application of Pharmacophores in Natural Product Research	128
3.1	Screening for Selective Inhibitors of 11 β -Hydroxysteroid Dehydrogenase 1	128
3.2	Identification of Novel Natural Inhibitors of <i>Trypanosoma brucei</i> Glyceraldehyde-3-Phosphate Dehydrogenase	132
	References	133

1 Introduction

In medicinal chemistry, the concept of pharmacophores has become increasingly popular over the last few decades and pharmacophore-based methods can be considered as an indispensable component in the modern computer-aided drug design toolbox. Due to their abstract nature, pharmacophores are easy to comprehend and

T. Seidel (✉) · A. Garon · T. Langer
Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria
e-mail: thomas.seidel@univie.ac.at; arthur.garon@univie.ac.at; thierry.langer@univie.ac.at

D. A. Schuetz
InteLigand GmbH, IRIC–Institut de Recherche en Immunologie et en Cancérologie, Université de Montréal, Montréal, QC, Canada
e-mail: doris.alexandra.schuetz@umontreal.ca

© Springer Nature Switzerland AG 2019

A. D. Kinghorn, H. Falk, S. Gibbons, J. Kobayashi, Y. Asakawa, J.-K. Liu (eds.),
Progress in the Chemistry of Organic Natural Products, Vol. 110,
https://doi.org/10.1007/978-3-030-14632-0_4

99

intuitive, which renders them rather useful as a tool to describe, explain, and visualize ligand–target binding modes.

Depending on background and context, the term pharmacophore was often attributed with different meanings. Historically, medicinal chemists used the term pharmacophore to vaguely denote common structural or functional elements of a set of compounds that are essential for activity toward a particular biological target. However, the official IUPAC definition for this term [1] is more specific and states: “A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supra-molecular interactions with a specific biological target structure and to trigger (or to block) its biological response.”

According to this definition, pharmacophores do not represent sets of particular functional groups (e.g., a primary amine or thioamide) or characteristic structural fragments (e.g., a pyrrolidine ring), but are an abstract description of stereoelectronic properties of molecules that are indispensable for energetically favorable ligand–target interactions. Molecules possessing similar pharmacophoric patterns can therefore be assumed to be recognized by the same binding site of a biological target and thus also show similar biological profiles [2].

2 The Pharmacophore Concept

2.1 *Historical Background*

The idea that drug molecules can act upon some receptor was put forward by Langley in 1878 [3] who named receptors “receptive substance” [4]. The word “receptor” itself was introduced later by Paul Ehrlich [5, 6]. Although several observations during the first half of the twentieth century supported the receptor concept [7], the selectivity of drug–target interactions was not then commonly recognized and accepted. A notable milestone in this regard was the drug Salvarsan discovered by Paul Ehrlich. Ehrlich intended to develop a chemical compound that could be used as a chemotherapeutic “magic bullet” against specific infectious organisms. After testing hundreds of candidate compounds, he eventually came across one that had enough potential to treat syphilis and trypanosomiasis.

The therapeutic effect was confirmed in clinical tests and Ehrlich’s discovery was officially announced in 1910. A side effect of Ehrlich’s discovery was the support of an assertion made by Emil Fischer in 1894. From his research results, Fischer reasoned that an enzyme and glucoside must fit together like lock and key in order to have a chemical effect on each other [8]. This “lock and key” concept is still in use today although it assumes a rigid receptor structure. Currently, it is commonly known that the receptor part is also flexible and, at least to some extent, can adapt to the structure of a bound ligand.

Long before computers became an integral part of drug design and optimization, simple pharmacophores were already described in the literature and became applied by medicinal chemists for the development of novel drugs. Thanks to a knowledge of the bond lengths and van der Waals radii, early structure–activity relationship considerations were possible starting in the 1940s and allowed the construction of simple two-dimensional model structures.

Notable within this respect is the recognition of the ability of *p*-aminobenzoic acid (PABA, a biological precursor of dihydrofolic acid) to reverse the bacteriostatic effect of *p*-aminobenzenesulfonamides. This finding led later to the formulation of the fundamentals of the theory of metabolite antagonism by Woods and Fildes [9, 10]. As shown in Fig. 1, PABA and the sulfonamides are isosteres and either the metabolite or its antagonist can attach to the critical area on the dihydrofolate reductase enzyme surface. If the latter occurs, the metabolic process is interrupted and, in the case of bacteria, multiplication is inhibited.

Another early achievement was the discovery of (*E*)-diethylstilbestrol. (*E*)-Diethylstilbestrol acts as an estrogenic agent, which is due to its similarity to estradiol [11] (Fig. 2). The fact that the estradiol conformation is not planar was known even at that time; however, the proposed model was a two-dimensional one.

Chiral and conformational effects were first included in drug design considerations starting in the early twentieth century which eventually led to a deeper understanding of the interdependencies between three-dimensional ligand structures and associated activities. Furthermore, it became obvious that the simple presence of

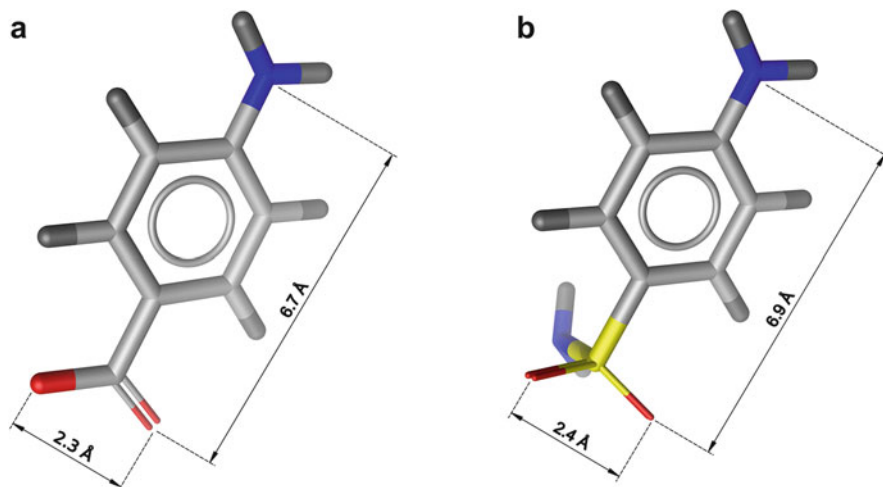


Fig. 1 *p*-Aminobenzoic acid (PABA, **a**) and *p*-aminobenzenesulfonamide (**b**) are isosteres and show similarities regarding interatomic distances that are critical for binding to the dihydrofolate reductase enzyme surface [2]. Binding of the sulfonamide instead of PABA thus inhibits the biosynthesis of tetrahydrofolic acid

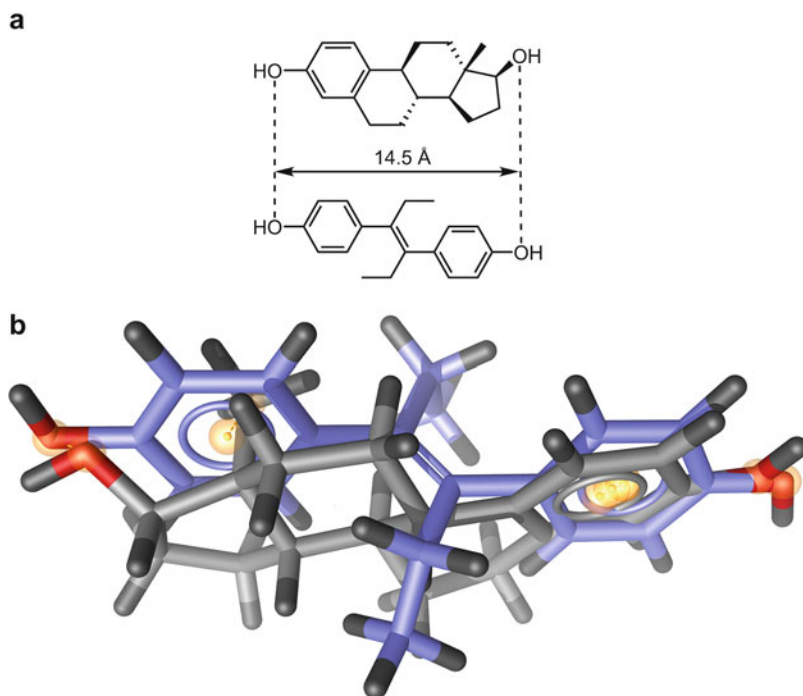


Fig. 2 Analogy between estradiol and (*E*)-diethylstilbestrol [2]: **(a)** 2D structures of both compounds, **(b)** three-dimensional overlay, the compounds with estradiol shown in gray and (*E*)-diethylstilbestrol in purple

pharmacophoric groups alone is not sufficient to explain biological activity and that the spatial disposition of these groups plays an equally important role in the recognition of a ligand by the target protein.

Easson and Stedman introduced the “three-point contact model” in 1933 [12], which suggested that the substituents of a chiral center establish a three-point contact with the target protein. This led to the assumption that only one enantiomer of the molecule can form a complementary match. The optical antipode of the active molecule will not present an adequate feature distribution. Underlining this, adrenaline (epinephrine) presents an interesting showcase. (*R*)-(-)-Adrenaline, which is the more active natural form, establishes contact with the adrenergic receptor by engaging a total of three interactions (Fig. 3); its stereoisomer (*S*)-(+)-adrenaline, which shows less activity, merely displays a two-point contact (Fig. 3). Consequently, the loss of an energetically favorable hydrogen bonding interaction leads to an approximately 100-fold lower activity of (*S*)-(+)-adrenaline when compared to (*R*)-(-)-adrenaline.

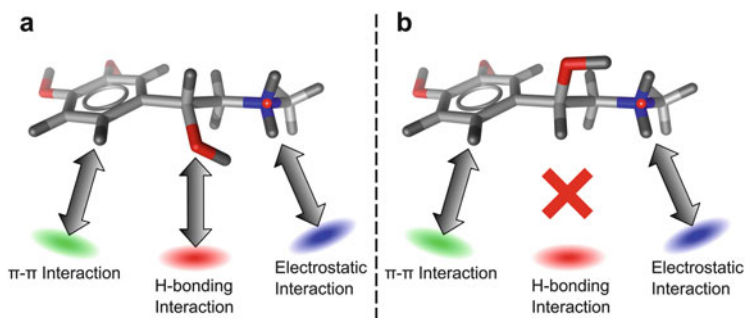


Fig. 3 Possible interactions of (*R*)-(-)-adrenaline (**a**) and its stereoisomer (*S*)-(+)-adrenaline (**b**) with an adrenergic target receptor [2]

The first crystal structures of protein targets [13] substantiated the drug–receptor theory. Being able to cocrystallize a drug bound to its target opened a new field of possibilities, as the interaction pattern of a small molecule and the protein can be examined in much more detail. Being able to analyze the interactions of the anticancer drug methotrexate bound to the dihydrofolate reductase enzyme was particularly revealing for the interplay of pharmacophoric ligand and receptor features [14]. With the availability of further protein structures, many more structural models could thus be generated based on their homology to solved crystal structures.

The next conceptual breakthrough was an answer to the question of whether chemical forces alone are sufficient to explain drug–receptor interactions and other pharmacological effects, or if yet unknown additional driving forces have also to be considered. Wolfenden performed a thorough investigation of intermolecular forces [15] which eventually led to a rationale for explaining drug–receptor interaction rates by chemical forces alone. This later evolved into present-day calculations of ligand–receptor binding energies, dynamic motion of ligand and receptor, and linear free-energy perturbation estimation.

The examples and milestones outlined above, together with influential work by Gund [13, 16], Humblet and Marshall [17], and many others, paved the way for the modern pharmacophore concept and its derived applications. The pharmacophore concept summarizes insight into the effects of chemical structure on bioactivity. This allows medicinal chemists to postulate pharmacophore models as the “essence” of the structure–activity knowledge they have gained in an extensive structural study of a series of active and inactive molecules for a given drug target.

2.2 Three-Dimensional Pharmacophores

Pharmacophores present a valuable tool to represent the nature as well as the location of functional groups of a molecule that is involved in ligand–target interactions. For an easy comprehension of such interactions by humans, all types of noncovalent

interactions can be represented by geometric entities. Apart from the abstract characterization of known structures, pharmacophore modeling can aid in the design of novel molecules and the prediction of their activities. One of the strengths of integrating the concept of pharmacophores is that it allows scaffold hopping, as the pharmacophoric features of a model have to be met in their spatial arrangement, but not necessarily the underlying chemical structure as such. These geometric entities, represented by spheres and vectors, prompt a simplified way to capture the conformation of a molecule and the features that contribute to its activity in three-dimensional space. Obviously, the choice of features (number and type) has an inevitable impact on the quality of a pharmacophore model.

While the choice of features used to be very specific at the early stage of pharmacophore modeling [18], recent techniques build pharmacophore models in a more general way [19]. Nonetheless, while those universal models are interpretable and easy to comprehend, they might lack selectivity. Building a very general model can sacrifice the quality of the pharmacophoric representation by reflecting chemical functionality only, but neglecting further characteristics of specific functional groups. However, building a very restrictive model by employing a higher number of feature types can quickly lead to problems when it comes to the identification of structurally unrelated chemical compounds. Thus, developing a feature set that represents a reasonable tradeoff between being too general and being too selective is one of the biggest challenges current software packages [20–23] for pharmacophore modeling face. In order to describe the different levels of universality and specificity of chemical features, a simple layer model according to Table 1 can be used [23–26]. In this model, a lower layer number corresponds to higher specificity and, therefore, lower universality. Some examples for chemical features together with the corresponding abstraction level are given in Table 1.

Table 1 Classification of the Abstraction Levels of Chemical Features

Layer	Example	Classification	Universality	Specificity
1	An ammonium group facing another aromatic system within a distance of 2–4 Å	Molecular graph descriptor (atom, bond) with geometric constraints	– –	+++
2	A hydroxy group, a primary, secondary, or tertiary amine, or a carboxylic acid moiety	Molecular graph descriptor (atom, bond) without geometric constraints	–	++
3	A hydrogen bond acceptor vector including the acceptor location as well as the projected donor point; an aromatic ring system with location and orientation (ring plane)	Chemical functionality (aromatic ring, hydrogen bond donor, acceptor) with geometric constraints	++	+
4	Hydrogen bond acceptor without a projected donor point; a lipophilic group	Chemical functionality (lipophilic area, positive ionizable group) without geometric constraints	+++	–

If a higher level definition (levels 3 and 4) does not describe sufficiently the features occurring in the training set, low universality levels 1 and 2 will be employed [27]. If such a customization results in a layer 1 or layer 2 feature, there should be a possibility of including layer 3 or 4 information in order to be able to categorize and thus increase comparability (e.g., an ammonium group as a layer 2 feature is a subcategory of “positive ionizable,” which is a layer 4 feature).

Some of the most important types of ligand–receptor interactions together with their corresponding representation in pharmacophore models are discussed in the following sections.

2.2.1 Basic Interactions and Their Representation

2.2.1.1 Hydrogen Bonding Interactions

Hydrogen bonding can usually be observed when an electropositive hydrogen atom interacts with a so-called hydrogen bond acceptor. A hydrogen bond acceptor—or in short “H-bond acceptor”—is an electronegative atom like oxygen, fluorine, or nitrogen. The counterpart, the H-bond donor, provides the hydrogen, which is covalently bound to another electronegative atom. The hydrogen bond can be created between those two atoms and it indeed represents the most important specific interaction observed in the formation of ligand–receptor complexes [28]. Hydrogen bond acceptors and donors are usually modeled as a position, allowing a certain tolerance to that position as well as for the position of its counterpart. Together these two positions form a vector that constrains and directs the H-bonding axis as well as the location of the interacting atom in the protein target. Therefore, donor and acceptor features, since they show specific directional constraints, are layer 3 features. If the direction constraint is omitted, they become layer 4 features. This makes them much less specific and they can match any acceptor/donor atom irrespective of whether the essential geometric preconditions for H-bond formation are fulfilled (Fig. 4).

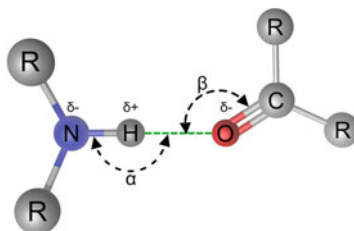


Fig. 4 Geometry of hydrogen bonding: The N and H atom of the secondary amine and the O atom of the ketone are aligned linearly. The distance between N and O is typically around 2.8 to 3.2 Å, while the N-H-O angle α is $>150^\circ$, the C=O-H angle β ranges from 100° to 180°

2.2.1.2 Hydrophobic Contacts

Aromatic and aliphatic hydrocarbons as well as halogens represent hydrophobic moieties of the ligand. Hydrophobic contacts result from the spatial vicinity of the side chain of a nonpolar amino acid and the aforementioned hydrophobic substituents on the ligand. Contribution to the binding free energy derives from an increase of entropy of the system, mainly because hydrophobic contacts normally lead to release of water molecules from hydrophobic areas (Fig. 5). Water molecules at hydrophobic interfaces display a very high degree of order without undergoing interactions and are often described as “trapped water molecules.” Yet, the unconstrained water molecules once released to the bulk solvent are capable of participating in energetically favorable hydrogen bond interactions and, therefore, contribute to a ligand’s overall binding affinity. According to $\Delta G = \Delta H - T\Delta S$, both contributions will lower the free energy change ΔG for the interaction, thus increase the ligand’s binding affinity.

Hydrophobic interactions are nondirectional, which enables their representation as unconstrained layer 4 features. A tolerance sphere is placed in the center of the hydrophobic moiety or chain.

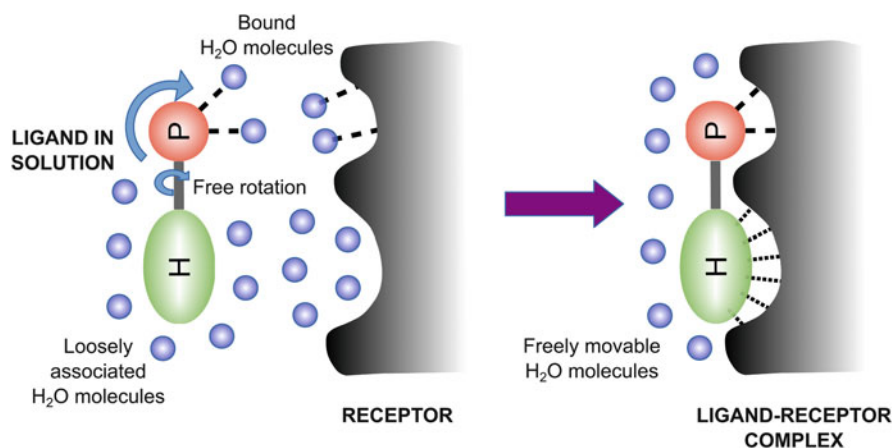


Fig. 5 On formation of interactions of hydrophobic ligand groups (H) and hydrophobic receptor structures, water molecules get expelled from the binding pocket. The desolvation of the binding pocket increases the entropy of the system, as the water molecules, which were trapped in an unfavorable (hydrophobic) environment, have been liberated. This contributes to an increase of the binding affinity of the ligand to its target, which is typically around -100 and -200 J/mol per \AA^2 of the hydrophobic contact surface [29]

2.2.1.3 Aromatic and Cation- π Interactions

These interactions are directional and strongly attractive, always involving at least one electron-rich aromatic moiety. The interactions can be found between π systems of two aromatic rings, which can also be called π stacking, and furthermore, between aromatic systems and nearby cationic groups like metal ions or ammonium cations of the protein side chains [30]. These interactions play a crucial role for stabilizing DNA and protein structure, as well as enzymatic catalysis. Interaction energies involving π systems are energetically comparable to hydrogen bonding and therefore contribute essentially to the binding free energy. Interactions of the π - π and cation- π types require specific geometric configuration of the interacting counterparts (Fig. 6). Like hydrophobic interactions, aromatic features in pharmacophore models are layer 4 features. They are represented by a tolerance sphere located in the center of the aromatic ring. Furthermore, the directional aspect of aromatic interactions is taken into account by additional information regarding the spatial orientation of the aromatic system. This can be done in the form of a ring plane or by two points defining this vector (layer 3 feature).

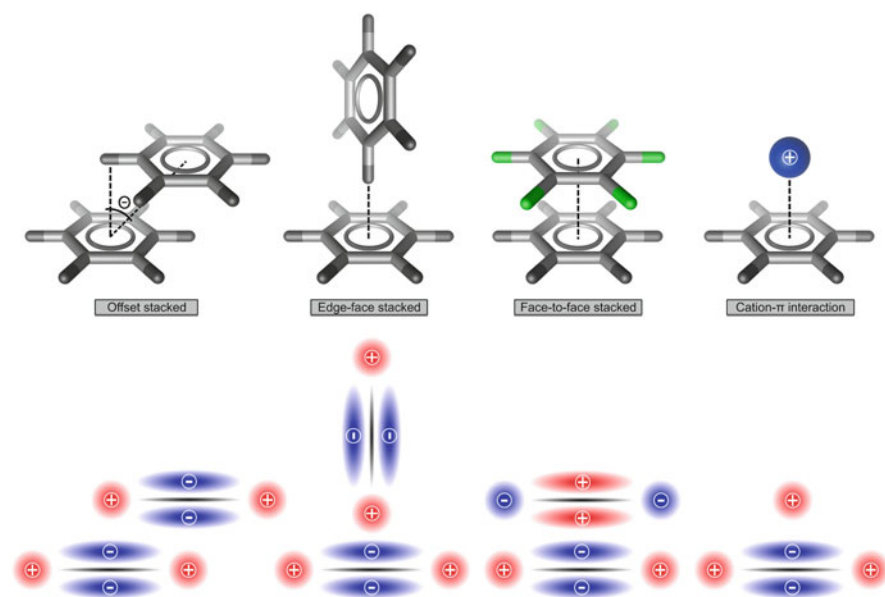


Fig. 6 Steric configurations of π - π and cation- π interactions [31]

2.2.1.4 Ionic Interactions

Oppositely charged groups attract each other and form relatively strong (>400 kJ/mol) interactions between the ligand and the protein environment. Those charged groups can be either single atoms like metal cations or functional groups that tend to easily protonate or deprotonate, such as carboxylic acids, guanidines, or aromatic heterocycles. These charged atoms or groups can be represented as positive or negative ionizable areas. Ionic interactions are nondirectional, as they are of electrostatic nature. In terms of pharmacophoric representation, this allows the utilization of simple tolerance spheres (layer 4 features).

2.2.1.5 Metal Ion Complexation

A number of proteins contain metal ions as cofactors. A prominent example of metal-containing enzymes is metalloproteases [32] which enclose Zn^{2+} , coordinated to the protein via three residues (Fig. 7). The complexation of this metal ion with suitable electron donating atoms or functional groups of the ligand can contribute crucially to the binding affinity and has proven essential for the ligand's mode of action. Functional groups that exhibit a strong affinity for metal ions are, for instance, thiols (R-SH), hydroxamates (R-CONHOH), or sulfur and nitrogen containing heterocycles. In pharmacophore models, metal ion binding interactions are usually represented by tolerance spheres located on single atoms or in the center of groups, capable of interacting with metal ions. To additionally constrain the location of the coordinated metal ion and/or to accommodate for a particular coordination geometry, a vector representation similar to hydrogen bonding interactions can be used.

2.2.1.6 Ligand Shape Constraints

A pharmacophore model does not necessarily represent all sufficient chemical characteristics an active molecule must exhibit for high-affinity binding. Even if a molecule fits the pharmacophoric model well and displays the features represented in the model, the molecule can still fail to bind due to steric clashes. The possibility of steric clashes between ligand and receptor atoms has to be considered by a model and can be handled by the incorporation of exclusion volumes. Such volumes can be of different sizes as they represent areas where the ligand is not allowed to occupy space after an alignment with the pharmacophore. The structures from an X-ray analysis of the receptor can be used to extract reliable information about spatial restrictions upon ligand-target binding. Utilizing a crystal structure enables the representation of exclusion volumes in accordance with the size of respective residues present in the receptor binding site. The van der Waals radii of the corresponding atoms determine the size of the exclusion spheres (Fig. 8), and a clash with those spheres coincides with the ligand atoms overlapping the receptor atoms. This naturally leads to a poor fit of the molecule inside the binding site and presumably to no ligand binding. In various cases, X-ray structures of the receptor

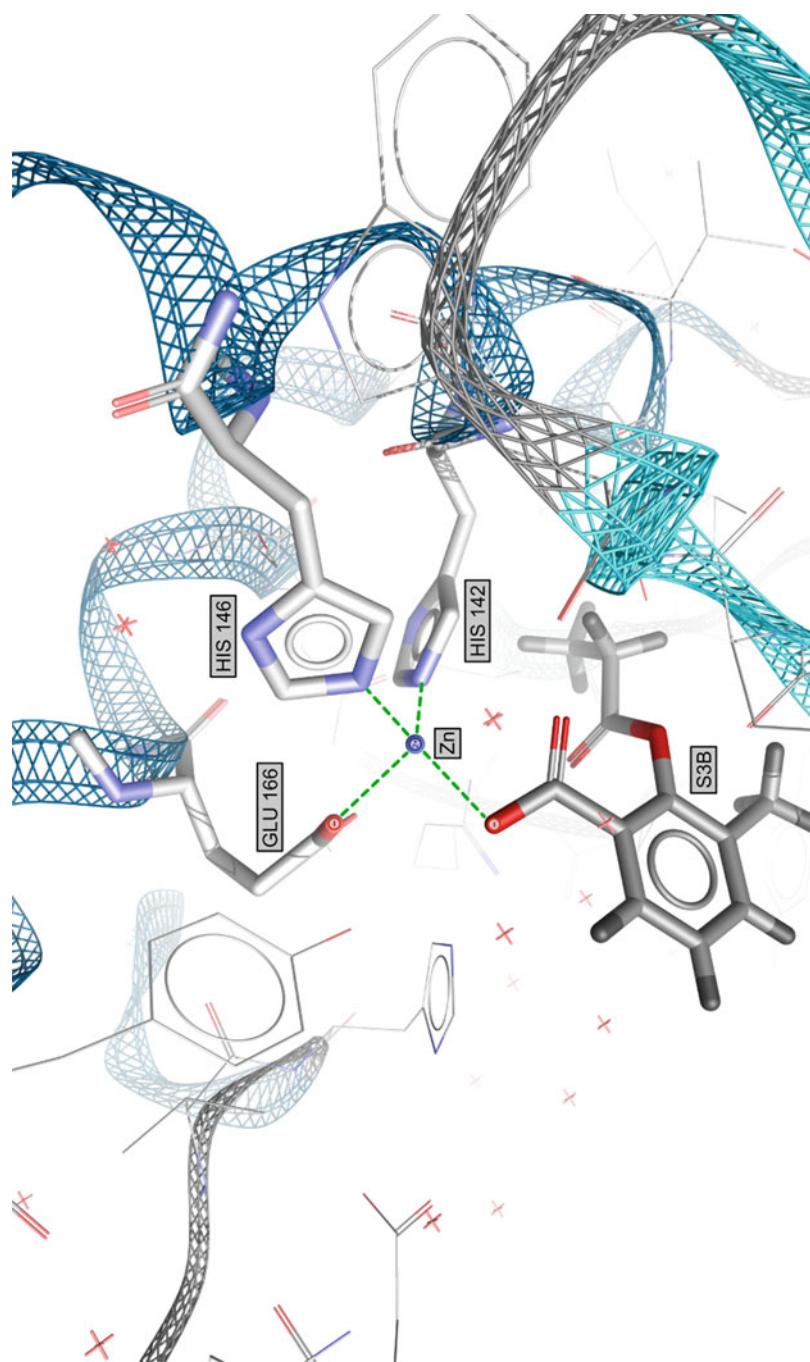


Fig. 7 Thermolysin in complex with the inhibitor 3-methyl-2-(propanoyloxy)benzoic acid (S3B, PDB-code: 3F2P). The tetra-coordinated Zn²⁺ ion interacts with the amino acids Glu166, Hist142, and His146 of thermolysin, and the carboxylate oxygen of the ligand [33]

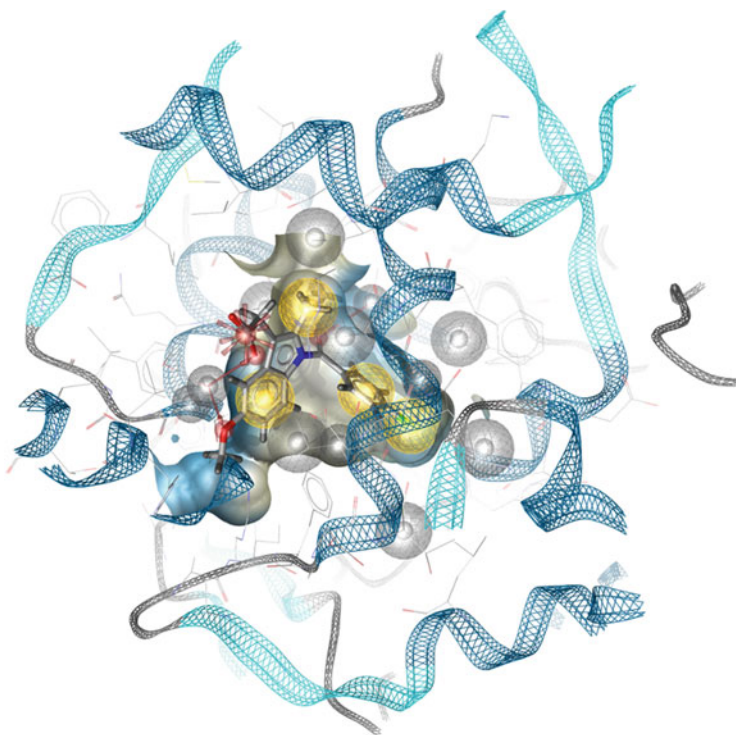


Fig. 8 Receptor-based pharmacophore generated by LigandScout for the COX2/inhibitor complex 4COX. Gray spheres represent exclusion volumes that model the shape of the receptor surface. Yellow spheres represent hydrophobic, red arrows hydrogen bond acceptor features, and the red spherical star represents a negative ionizable group involved in an electrostatic interaction

are not available, which makes placing volume spheres more challenging. The location and size of the exclusion volumes must be assigned manually or computer-aided methods can be used to distribute the volume spheres based on the union molecular shapes of a set of aligned known actives.

2.2.2 Pharmacophore Elucidation

Pharmacophore models can be constructed manually, and may be generated in an automated way starting from the structure of one or multiple ligands (ligand-based), or they can be deduced from the three-dimensional structure of the target receptor (structure-based). Which approach is chosen depends mainly on data availability, data quality, and computational resources. It also plays a crucial role for which purpose the pharmacophore model will be used. The most common approaches for pharmacophore modeling and their characteristics are outlined in the following sections.

2.2.2.1 Manually Created Pharmacophore Models

From an algorithmic point of view, this is the simplest way to obtain pharmacophore models. Manual construction relies on information about known key characteristics and/or the molecular structures of a series of active compounds. Choosing a manual construction can be feasible if an experimental structure of the bound ligand is available or the ligand exhibits only very low conformational flexibility. This is because one of the biggest sources of uncertainty in pharmacophore modeling is conformational flexibility, which makes a determination of optimal feature positions challenging. Placing pharmacophoric features for a rigid compound or for a known active ligand conformation is thus already much less complicated. Another challenge is to choose the set of relevant features that needs to be incorporated into the pharmacophore model. Today, this task is normally left to computer-aided methods. The manual involvement has moved toward the refinement of the model, while the model itself is automatically generated using specialized software tools.

2.2.2.2 Receptor-Based Pharmacophore Models

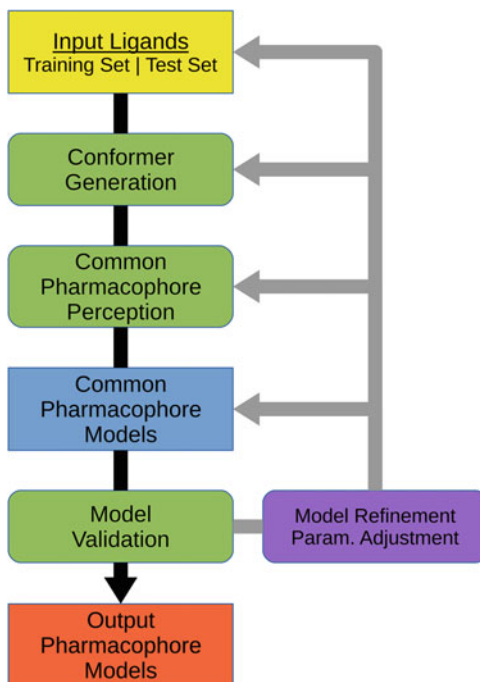
Having access to three-dimensional information about a ligand–receptor complex offers a tremendous advantage for developing high-quality pharmacophore models. The bioactive conformation of the ligand can be retrieved directly from the atomic coordinates set of the investigated complex and guides the correct placement of the pharmacophoric features. Information about the binding site structure moreover enables the incorporation of receptor shape information into the generated models (see Sect. 2.2.1). Potential interaction points and interactions of interest are identified by a thorough binding site analysis using an array of available methods [34]. Grid-based methods like GRID [35, 36] probe the binding site at discrete points employing small molecules or functional groups. Interaction energies between probe molecules and the receptor atoms are calculated, which finally results in so-called molecular interaction fields (MIF). These fields can then be used to identify energetically favorable and unfavorable regions for specific ligand–receptor interactions that guide a correct pharmacophoric feature placement or aids in ligand design and optimization. The program GRAIL [37], a quite similar but pharmacophore-based interaction field method, can also be employed for the prediction of interaction hot spots in apo binding sites (see Sect. 2.4.2 for a more thorough description of GRID and GRAIL). The programs LUDI [38] and SuperStar [39] are further available methods, which use a knowledge-based approach for the identification of interaction hot spots in the binding site. The rules these methods rely on are largely based on a statistical interaction analysis of experimental structures. They take into consideration the chemical nature of the functional groups involved as well as the preferred orientation of directed interactions, as is the case for hydrogen bond donors and acceptors. After the identification of all possible interaction sites, three-dimensional pharmacophores with three or more features (at least three features are required for a defined orientation in space) can then be generated.

These models have to be validated thoroughly and any models with no significance get discarded. This can be done, for example, using enrichment-based methods. Commercially available programs that are able to perform the entire modeling process from structure to pharmacophore include structure-based focusing [40, 41] and LigandScout [24, 25]. In structure-based focusing [42], a sphere with user-adjustable location and size is used to mark key residues in the binding site and a LUDI interaction map is generated to identify favorable interactions in which a ligand is expected to engage. The interaction map is transferred to an interaction model, which consists of a set of complementary points in the binding pocket, representing possible locations of pharmacophore features on the ligand. A user-defined density controls the number of points created, but it is usually quite large. Therefore, hierarchical clustering is performed to select a smaller number of representative features. After the addition of exclusion volumes, a database of known actives is searched to determine which pharmacophore models are most frequently matched. LigandScout takes a more direct approach and derives a pharmacophore model from a single ligand/receptor complex. After perceiving hybridization states, unsaturated bonds, and aromatic rings, the ligand and binding pocket structure is analyzed for the presence of atoms and groups that can take part in hydrogen bonding, hydrophobic, aromatic, ionic, and metal binding interactions. Pharmacophoric feature detection can be customized with respect to interaction-specific geometric characteristics like allowed distances and angle ranges. Whether a feature is incorporated into the final pharmacophore model depends on its location relative to a complementary feature in the binding site. For example, a hydrogen bond acceptor feature of the ligand is only included if there is an opposing hydrogen donor feature on the receptor side within a certain distance and angle range. After all complementary feature pairs of the complex have been detected and the corresponding ligand side features have been put into the derived pharmacophore model, exclusion volume spheres are finally added to resemble the shape of the binding pocket. Figure 8 shows a typical receptor-based pharmacophore that was generated by LigandScout for the cyclooxygenase-2 complex 4COX.

2.2.2.3 Ligand-Based Pharmacophore Models

Ligand-based pharmacophore modeling is the way to go in cases where no information about the three-dimensional structure of the receptor is available, but a sufficient number of actives are known. A very important prerequisite for developing a correct model of high quality is knowledge about the binding mode of the ligands. The ligands from which the pharmacophore model is derived must bind to the same receptor, at the same binding site, and in the same orientation. Otherwise, the pharmacophore models obtained will not represent the correct mode of action and cannot be used for the identification of novel actives via, for example, virtual screening experiments. For the generation of pharmacophore models starting from a set of active ligands, several algorithms have been published that are described elsewhere [34, 43]. In general, they all follow the generic workflow shown in Fig. 9.

Fig. 9 Workflow for the generation of ligand-based pharmacophore models. The input structures are known, active compounds. For these structures, multiple conformations will be generated (if not provided as an input) and common pharmacophore(s) will be perceived. Finally, a pharmacophore validation procedure provides information about the quality of the models obtained. If necessary, the models will then be refined or completely rebuilt using different training/test set molecules and/or by changing the parameters of previous processing steps



The input structures are imported and prepared accordingly. After this, conformers are generated. This step is a delicate and crucial part of the procedure, as the conformer output has to provide a dataset that is sufficiently large and diverse. This step is so important, as the bioactive conformation of the ligands is not known. Therefore, more than one conformation has to be generated and in each set of generated conformers at least one of the structures should represent a good approximation of the bioactive conformation. In the next step, a chemical feature pattern has to be identified [26], which is common in all training set ligands, and furthermore, can be superimposed with one or more conformations of each ligand. In many cases, more than one pharmacophoric pattern can be extracted. This results in a list of multiple possible solutions that are ranked according to a fitness function. The best model(s) are usually selected from the list by the user, following a careful validation procedure [44]. A validation of pharmacophore models can be carried out in various ways [43]:

1. Statistical significance analysis and randomization tests.
2. Enrichment-based methods: the ability to recover active molecules from a test database by placing a small number of known actives among randomly selected compounds. Pharmacophore-based virtual screening techniques and receiver operating characteristic (ROC) curve plots are usually employed for this validation approach (see also Sect. 2.3.1).
3. Biological testing of the retrieved, putatively active molecule.

Manual refinement should be considered if the pharmacophore validation procedure reveals an unsatisfying quality of the generated model. This can be done by a deletion/addition of features or by changing feature tolerances. However, the entire modeling procedure can also be repeated using a different setup. The composition of the training and test set can be altered, conformer generation might be repeated with fine-tuned parameters, and of course the pharmacophore model generation itself can be carried out using different settings. Ligand-based pharmacophore modeling is challenging, as a number of variables impact the outcome and the receptor structure is completely ignored. It is crucial to keep a critical view on the models and to validate comprehensively. Furthermore, the expertise of the user and the algorithmic power of the software employed are contributing significantly to a successful ligand-based pharmacophore model.

2.3 *Application of Pharmacophores in Drug Design*

2.3.1 **Virtual Screening**

The most common reason for the generation of pharmacophore models is their use as an efficient search filter for the discovery of novel compounds with a set of desired stereoelectronic properties. Most chemists are familiar with two-dimensional substructure-based similarity searches, but these typically yield hit lists containing only compounds of the same structural family or with largely similar scaffolds. Pharmacophore models, though, represent an abstraction that is able to also identify alternative chemotypes, that is, compounds with a different underlying common framework or scaffold. All that matters for a match to the pharmacophore query is the spatial disposition of stereoelectronic molecular features, and not the underlying chemical structure on the atom and bond level. The simplicity of the pharmacophoric abstraction enables a fast *in silico* search even of large compound databases. Molecules that end up in the obtained hit lists are guaranteed to exhibit the desired pharmacophoric features and thus have a high chance to show biological activity toward the target of interest [45–48]. Depending on the selectivity of the query pharmacophore, application-specific match constraints, and database size, tens to thousands of hit molecules are usually retrieved by a typical screening run. A portion of the hits will be false positives and show no significant activity at all, but in comparison with random sampling, hit rates obtained by pharmacophore searches are generally much higher and thus allow an enrichment of potentially active compounds in significantly smaller database subsets. This is especially important for academic research where the available resources are usually quite limited and only budget and facilities for a low-throughput biological testing of relatively few compounds might be at hand. In industry, on the other hand, it is often equally cost-effective to screen the whole corporate collection as it is to screen just a significant subset [34]. The main role of industrial pharmacophore searching thus has changed and the emphasis is nowadays on the creation of small focused sets for low-throughput, higher quality assays, which are carried out in parallel with high-

throughput screening experiments in order to enhance the lead identification process. As already stated, pharmacophore screening allows the identification of novel and diverse chemotypes, which are for a human researcher not obvious matches to the query pharmacophore. Hit lists containing molecules that belong to different structural classes thus can also serve as a valuable source of new “ideas” for the development and optimization of novel lead compounds that might not have been discovered by traditional rational drug design processes alone.

The next sections provide some insight into the algorithmic details of pharmacophore-based database searching and the preparatory work usually required. After this, the chapter concludes by a discussion of the methods and measures that can be used for hit-list analysis and the assessment of the discriminatory power of the query pharmacophores.

2.3.1.1 Annotated Database Creation

An important aspect to consider when screening compound collections against three-dimensional pharmacophore models is conformational flexibility. The most common way to deal with this problem is to create dedicated screening databases that store pregenerated conformations for each of the contained molecules. Another approach is to tweak the conformation of the molecules on-the-fly in the pharmacophore matching process [49]. The latter approach has the advantage of lower database storage requirements; however, screening runs computationally are more expensive and thus considerably slower. An additional disadvantage is the dramatic reduction of the conformational search space during pharmacophore alignment that bears the danger of getting trapped in a local minimum [34, 50]. Currently, pregenerating conformations and storing them in dedicated databases is the preferred approach, as hard disk storage is cheap and readily available. Furthermore, the generation of screening databases has to be performed only once, as they can be reused for all subsequent screening runs. This obviously results in a considerable acceleration of the overall screening process and allows to screen even millions of compounds in relatively short periods of time.

2.3.1.2 Database Searching

The database search is most commonly implemented as a multistep filtering process. First, a fast prefiltering step based on feature types, feature counts, and quick distance checks is applied in which definitively nonmatching compounds can get eliminated to an already large extent. Second, exact 3D matching algorithms are utilized, which are normally slower but more restrictive.

2.3.1.3 Prefiltering

Since the actual three-dimensional alignment of query pharmacophore models and molecules is the time-limiting step in the screening process, prefiltering is of utmost importance [51]. Prefiltering aims at a quick identification and elimination of all molecular structures that cannot be fitted to the query pharmacophore model in three dimensions. Only molecules that pass this filtering step need to be processed in the final, accurate, but computationally expensive three-dimensional alignment step. Descriptor-based similarity methods [52] have proven to be appropriate filters, as little information is needed, and similarity calculations are fast and the implication of biological similarity from structural similarity is generally valid [53].

Feature count matching is a very simple, but nevertheless effective filtering method that is able to dispose a large fraction of the database molecules (depending on the complexity of the query) in a computationally inexpensive way [34]. If feature counts are determined for the query pharmacophore model and get precalculated for the database molecules, only molecules that have the same (or higher) feature counts as the query need to be forwarded to the time-consuming matching step. Another medium complex method in terms of three-dimensional pharmacophore database searching is the concept of “pharmacophore keys” [54]. Essentially, pharmacophore keys are simple binary fingerprints that encode the spatial disposition of features in three-dimensional pharmacophores. By a binning of interfeature distances and the calculation of hash codes for all possible two-, three-, or four-point feature subsets of an input pharmacophore, a set of specific bit indices in a fixed size bitset is obtained in which each bit then denotes the presence or absence of a particular n-point pharmacophore. As a result, the screening becomes a simple intersection test to identify molecules that do not satisfy the query. Nearly all currently available software applications use similar approaches that follow this concept with only slight modifications like integration of feature tolerance sampling, different feature definitions, varying binning constraints, and so on [55, 56]. Most programs also include filters that potentially discard molecules that mathematically could fit the query pharmacophore model, but this loss in filtering is accepted for the benefit of higher efficiency. Other programs, such as LigandScout, strictly apply lossless filters that guarantee all of the discarded molecules are not able to geometrically match the query, which results in geometrically more accurate virtual screening results [57]. In summary, prefiltering is intended to prune large parts of the overall search space in favor of speed. This speed-up strategy, however, must ensure that the overall quality of the screening outcome is maintained and goals like an enrichment of actives and the identification of novel scaffolds [58] can still be met.

2.3.1.4 Matching of Three-Dimensional Pharmacophore Models

Once all database molecules that have a high chance to match the query pharmacophore are identified, their conformation-specific pharmacophores need to be examined more closely to see whether they are able to match the spatial disposition of the query features. In this step, it will finally be decided whether a database

compound gets rejected or is put into the final hit list. In general, special care must be taken in this decision process since it has a direct impact on the quality of the screening results obtained. The geometric alignment of a query pharmacophore model with the pharmacophore derived from a database molecule conformation can be reduced to the problem of finding a suitable subset of features that fulfills all n -point distance combinations of the query. Greedy algorithms that find solutions for this problem have been proposed relatively early and range from three-dimensional maximum clique detection algorithms [59] to the incremental buildup of increasingly larger common feature configurations [60]. However, since pure feature-pair distance comparisons (two-point pharmacophores) cannot distinguish between a pharmacophore and its mirror image [19], an actual overlay in three-dimensional space is required to be able to correctly identify a match to the query within the defined feature tolerances. This overlay is also necessary to check and/or score additional constraints imposed by vector features like hydrogen bond acceptors/donors, planary features like aromatic rings and exclusion/inclusion volume spheres. Commercial software packages for pharmacophore modeling that incorporate state-of-the-art screening functionality like catalyst [20], phase [56], MOE [22], and LigandScout [23] all perform some sort of geometric alignment in the three-dimensional pharmacophore matching step, which is usually done by minimizing the RMSD between associated feature pairs [61]. While all other programs implement a search in increasing n -point distances, LigandScout uses a sophisticated pattern-matching technique to identify an initial alignment resulting in lower restrictions regarding the number of features in the query pharmacophore model. Although the general strategies for hit identification are similar, they differ in various details that range from the handling of conformational flexibility and interpretation of query feature constraints to the customization of search parameters [62].

2.3.1.5 Hit-List Analysis

The hit list obtained by a database is a good starting point for the validation and refinement of the pharmacophore model (Fig. 10). For this purpose, several useful measures [51, 63–65] have been devised that are described in more detail below.

Sensitivity (Se) is the ratio of the retrieved true-positive compounds TP (Fig. 10) to all active compounds in the database, which is the sum of TP and the number of false-negative compounds FN . Sensitivity values can range from 0 to 1, where $Se = 0$ means that the search did not find any of the actives in the database and $Se = 1$ means that the search returned all active compounds.

$$Se = \frac{TP}{TP + FN}$$

Specificity (Sp) is the amount of rejected truly negative compounds TN divided by the sum of TN and the number of retrieved false-positive compounds FP . Specificity ranges from 0 to 1 and denotes the percentage of truly inactive compounds. A value of $Sp = 0$ means that none of the inactive compounds could be

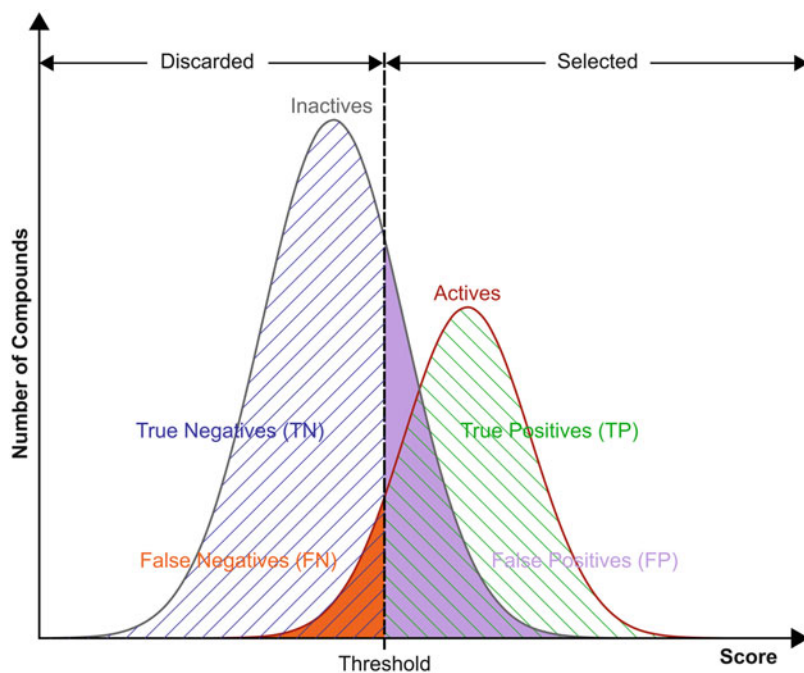


Fig. 10 Binary classification of compounds in a screening database into active and inactive relative to a set score threshold. Depending on the score distributions of the known active and inactive compounds (in gray and red) and the value of the score threshold, the classified compounds can be divided into four different subsets: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)

identified as such and $Sp = 1$ means that all inactive compounds have been correctly rejected during the screening process.

$$Sp = \frac{TN}{TN + FP}$$

Yield of actives (Ya) is a measure that shows the amount of the retrieved truly active compounds TP in relation to the size of the hit list n . The yield of actives can, for example, be used to compare hit lists retrieved for databases created with different conformer sampling techniques [64].

$$Ya = \frac{TP}{n}$$

Enrichment factor (EF) measures the yield of actives proportionally to the ratio of actives in the database, where A is the amount of actives in the database and N is the total number of database molecules (not including their conformations).

$$EF = \frac{Y_a}{A/N}$$

Goodness of hit list (*GH*) combines sensitivity, specificity, and yield of actives and is therefore a very useful measure that considers both the true actives ratio and the true inactives ratio. The goodness of hit list is defined as the weighted sum of Y_a and Se multiplied with Sp . The quantity of active compounds is usually weighted higher than that of actives in the hit list. For example, Güner and Henry [63] weight the yield of actives with 3/4 and the sensitivity with only 1/4. Thus, a high value of *GH* can only be achieved with a high value of actives and a low false-negative ratio at the same time.

$$GH = (w_1 \cdot Y_a + w_2 \cdot Se) \cdot Sp$$

A modern tool for the assessment of screening results is a receiver operating characteristic (ROC) curve [66, 67] (Fig. 11). The ROC curve displays the increase

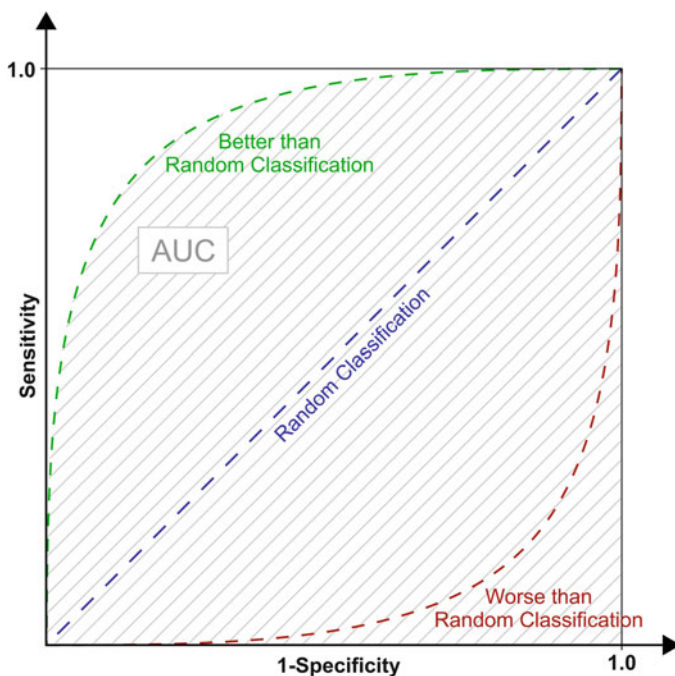


Fig. 11 Example of a ROC curve. The best possible classification method would yield a point in the upper left corner of the ROC space, representing 100% sensitivity (i.e., no false negatives) and 100% specificity (i.e., no false positives). Random guesses would result in points along the diagonal line (blue) from the left bottom to the top right corners. Points above the diagonal represent good classification results (green; better than random), points below the line represent bad results (red; worse than random). Note that the output of a consistently bad predictor could simply be inverted to obtain a good predictor. The classification model performance is determined by looking at the area under the ROC curve (AUC). The best possible AUC is 1, while the worst is 0.5. AUC values less than 0.5 suggest that one can simply do the exact opposite of what the model recommends to get back a value above 0.5

of false positives that results with increased true positives. The y -coordinate of the ROC curve represents the true-positive rate (sensitivity), whereas the x -coordinate denotes the appropriate false-positive rate (1 -specificity). An ideal curve would rise vertically along the y -axis until it reaches the maximum true-positive rate, which is 1 , and then continues horizontally to the right, which means that the hit list contains all active compounds in the database and that none of the hits is a false positive. The ROC curve of a random database search is represented by a diagonal line.

2.3.1.6 Pharmacophore Model Refinement

On the basis of an analysis of the hit list with the above measures and tools, the pharmacophore model is often refined to achieve more satisfactory results. Adaptation of feature definitions, modification of feature tolerances, addition or removal of features, and exclusion volumes are some of the adjustments that can help to tune a pharmacophore model.

Another possibility is to modify the database by readjusting the number of pregenerated conformations to address molecular flexibility more adequately. Since pharmacophore modeling and database screening are very complex tasks, several iterations of screening—analysis—refinement are usually necessary to achieve good results.

2.3.2 Pharmacophore-Based De Novo Design

Another application of three-dimensional pharmacophores worth mentioning is the tailored de novo design of ligands for a given biological target of interest. The hit compounds obtained from pharmacophore-based virtual screening runs are usually existing compounds that might be patent protected, show only unsatisfying activity, or have suboptimal ADME-Tox properties. In contrast to virtual screening, de novo ligand design allows the generation of novel molecular structures with desired pharmacological properties directly from scratch. In this approach, the de novo molecular structure generator is confronted with a huge search space of chemically feasible drug-like molecules that has been estimated to be in the order of 10^{60} – 10^{100} [68]. Such a large space renders an exhaustive search impossible. Instead of the systematic generation and assessment of each individual compound, a practically feasible de novo design process has to rely on the principle of local optimization which eventually ends up with some practical optimum solution for the given design problem. However, it is important to note that de novo design will rarely yield novel lead compounds with an already optimal activity in the nanomolar range. The molecules generated rather represent suggestions for a new series of lead structures

showing only micromolar activity and further (manual) optimization of the obtained compounds might be necessary.

An additional considerable challenge for de novo structure generators is to assess the synthetic accessibility of the designed compounds. This needs to be done because the generated molecules usually do not yet physically exist and have to be synthesized for more thorough investigations. A quite complicated or impossible synthesis of the generated molecules renders them practically useless and the generation of such structures must be avoided at all costs.

Up to now a considerable amount of de novo design methods have already been developed. These include, for example, 3D Skeletons [69], LEGEND [70], LUDI [38], NEWLEAD [71], CONCEPTS [72], SPROUT [73], MCSS&HOOK [74], SMoG [75], CONCERTS [76], LEA [77], LigBuilder [78], TOPAS [79], F-DycoBlock [80], ADAPT [81], SYNOPSIS [82], CoG [83], BREED [84], and PhDD [85] (for an excellent review of most of these methods, see [68]). All listed methods were applied in drug design projects and some of them also have shown good performance. However, most of the methods—with the exception of NEWLEAD and PhDD—adopt a strategy based purely on receptor structure. This means that detailed structural information about the target receptors must be available otherwise those methods simply cannot be applied. If the three-dimensional structure of the biological target is not available but one or more active ligands are known, a ligand-based strategy can be applied. Using pharmacophores to guide for the construction of novel molecular structures is especially appealing because they can be obtained both in a ligand- and receptor-based manner (see Sect. 2.2.2) and capture all essential ligand–target interactions independent of the underlying molecular structures.

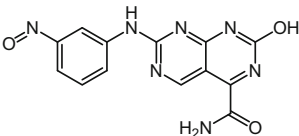
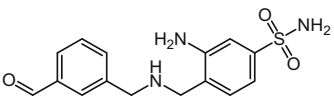
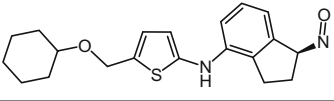
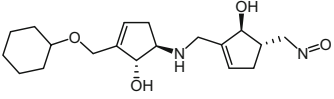
The program NEWLEAD, which was developed by Tschinke and Cohen in 1993 [71], was the first de novo design method that automatically generated candidate structures conforming to the requirements of a given pharmacophore. The program NEWLEAD builds novel structures by connecting a set of molecular fragments that correspond to key pharmacophoric features with spacers assembled from small chemical entities (atoms, chains, or ring moieties). The method was tested on several sets of input fragments, each consisting of selected functional groups that were obtained from the bioactive conformation of reference molecules comprising methotrexate, indomethacin, and the HIV-1 protease inhibitor A74704. The program was not only able to reproduce the reference structures, it also produced additional meaningful structures that were chemically unrelated to the reference molecules and thus demonstrated its potential for de novo drug design.

The program PhDD, a more recent development introduced by Huang et al. in 2010 [85], is in many respects similar to NEWLEAD, but addresses additional issues in order to increase the quality of the design results obtained. The program PhDD has the following characteristics that are distinct from the commonly used receptor-based de novo design methods and NEWLEAD: (a) PhDD completely works on an abstract pharmacophore model level and does not require any prealigned molecular fragments as input. Furthermore, PhDD regards exclusion volumes and thus allows constraintment of the size and shape of the generated molecules. (b) PhDD assesses

the synthetic accessibility of the generated molecules. (c) Fragments and linkers that are used for molecule buildup were all obtained by fragmentation of existing drug molecules. This drastically reduces the molecular search space and increases drug likeness of the generated structures. (d) The bioactivity of designed molecules is estimated by using a fit value, which describes how well a ligand is aligned with the input pharmacophore model. The program PhDD was validated using three test pharmacophore models that were generated for sets of known HDAC, CDK2, and IN inhibitors. The results have shown that PhDD was indeed able to generate completely new molecules that were structurally different from any known inhibitors but still had high synthetic accessibility, high drug likeness, and matched the pharmacophore hypotheses quite well. Table 2 lists the structures of four compounds that were generated for the CDK2 pharmacophore hypothesis together with their molecular weight, fit, and synthetic accessibility scores.

Although promising results could be obtained, pharmacophore-based de novo design can be inherently problematic due to the nature of the underlying pharmacophoric abstraction. If, for example, the input pharmacophore describes the binding mode of actives inadequately, any generated molecules will likewise be far from optimal in terms of their activity. Furthermore, exclusion volumes only allow an approximative modeling of receptor shape. As a consequence, structures might be generated for which the atoms will collide with atoms of the target binding site. A third problematic factor is the bioactivity scoring of the generated compounds. For activity prediction only pharmacophoric features are available that do not provide enough information for a reliable calculation of binding affinities. They just can be used to calculate a geometric fit of the designed ligand and the input

Table 2 Examples for De Novo Compounds That Were Generated by PhDD [85] for a Given CDK2 Inhibitor Pharmacophore

Designed compound	Molecular weight (g/mol)	Fit score	Synthetic accessibility
	311	8.327	603
	319	8.193	738
	356	8.563	778
	350	8.797	781

pharmacophore model, which does not necessarily correlate well with actual interaction energies. All of these factors are challenges that need to be embraced by pharmacophore-based de novo design methods in order to promote further developments in this promising field.

2.4 *Current Research and Developments*

2.4.1 **Dynamic Pharmacophore Modeling and Virtual Screening**

As already discussed in Sect. 2.2.2, the availability of three-dimensional structural information about a ligand–receptor complex in general allows one to derive pharmacophore models with highest possible quality. Three-dimensional structures of ligand–receptor complexes are usually obtained via X-ray crystallography or NMR methods. However, such experimentally determined structures of a ligand–target complex represent just a single snapshot of a dynamic system and do not provide information about the conformational flexibility and dynamics of the ligand and residues constituting the binding pocket [86, 87]. A structure-based pharmacophore model derived from only a single set of three-dimensional coordinates obtained from, for example, an X-ray structure might therefore include or lack features as a result of the crystal specific packing of the ligand–receptor complex. One approach to overcome this problem is to use multiple crystal structures of active ligands that are complexed with the same target. The interactions present for each ligand are identified and then merged into a more refined pharmacophore hypothesis [88–90]. This approach, however, is limited to ligand–target complexes for which multiple crystal structures are available and it has to be assured that each ligand exhibits the same binding mode. Another way to obtain several valid atomic coordinates sets that is not limited to the availability of multiple experimental structures is to perform molecular dynamics (MD) simulations of the investigated systems. The thus obtained very large amount of information can then be exploited for pharmacophore modeling and pharmacophore-based virtual screening in several ways. Choudhury et al., for example, derived a pharmacophore model for each output conformation and then performed a ranking based on docking and screening results [91]. A downside of this approach is that it requires activity data for validating/ranking the pharmacophore models. Other methods rely on a clustering or averaging of the ligand–target complex MD trajectory to obtain representative coordinates sets or perform similar operations on the derived pharmacophore models [92–96]. A general problem with the aforementioned methods is that they do not automatically deliver “better” pharmacophores. They all require the intervention of an experienced human operator to analyze the intermediary results obtained and then to decide about any further steps. Furthermore, infrequently observed metastable protein conformations might actually be those that enable the energetically most beneficial ligand interactions [97, 98]. A blind selection of highly populated conformer or pharmacophore clusters thus does not necessarily result in pharmacophore

models of higher quality. In a recent publication, Wieder et al. analyzed the stability of individual pharmacophoric features that were observed during MD simulations of ligand–protein complexes [99]. A merged pharmacophore model was constructed that consisted of all observed features and their occurrence count was used for prioritization. This approach gave interesting insights into the dynamics of the pharmacophore models but led to problems when the models were subjected to virtual screening runs. In several cases the combination of features from different protein conformations resulted in models that actually could not be observed in any of the frames saved during the MD simulation. Furthermore, the selection of representative feature coordinates also caused substantial difficulties. In summary, the authors could not find general solutions for the problems encountered without again having to rely on the expertise of a human operator that performs model evaluation and refinement. These examples show that an improvement of pharmacophore hypotheses using MD conformational flexibility information is not as straightforward as it might seem at the first glance.

However, when the main focus of interest lies on the improvement of virtual screening results the picture changes completely. The common hits approach (CHA) [100], for example, is a fully automated procedure for pharmacophore-based virtual screening that aims at boosting early enrichment by making use of conformational flexibility information in a traditional screening protocol setting. The approach generates a structure-based pharmacophore model for each coordinates set that was saved during an extensive MD simulation and then groups all models together that have the same set of ligand side features. This step effectively reduces the high number of pharmacophore models obtained for all frames of the MD trajectory to only a few hundred representative models (RPM). Virtual screening runs are performed with each representative pharmacophore model and the screening results obtained are combined and rescored to generate a single hit list. The final score for a particular molecule in the output hit list is then calculated based on the number of representative pharmacophore models that matched the molecule (see Fig. 12 for a graphical representation of the CHA workflow).

The performance of the CHA approach was assessed using screening databases with actives and decoys for 40 protein–ligand systems. The assessment was performed by a comparison of the ROC AUC value obtained for the CHA hit list and the corresponding AUC values obtained for the pharmacophore model of the experimental PDB structure and a representative model of the most populated MD derived pharmacophore cluster. For 34 of the 40 investigated systems, for which at least one of the performed screening runs gave results better than a random classifier, in 68% of the cases the highest enrichment was achieved by the CHA, compared to 12% for the PDB structure model and 20% for the representative pharmacophore model of the most populated cluster. These results clearly show that an incorporation of conformational flexibility information can indeed help to increase the quality of results obtained by classical pharmacophore-based drug design techniques and a MD simulation of the investigated biological target might be worth the relatively high computational costs.

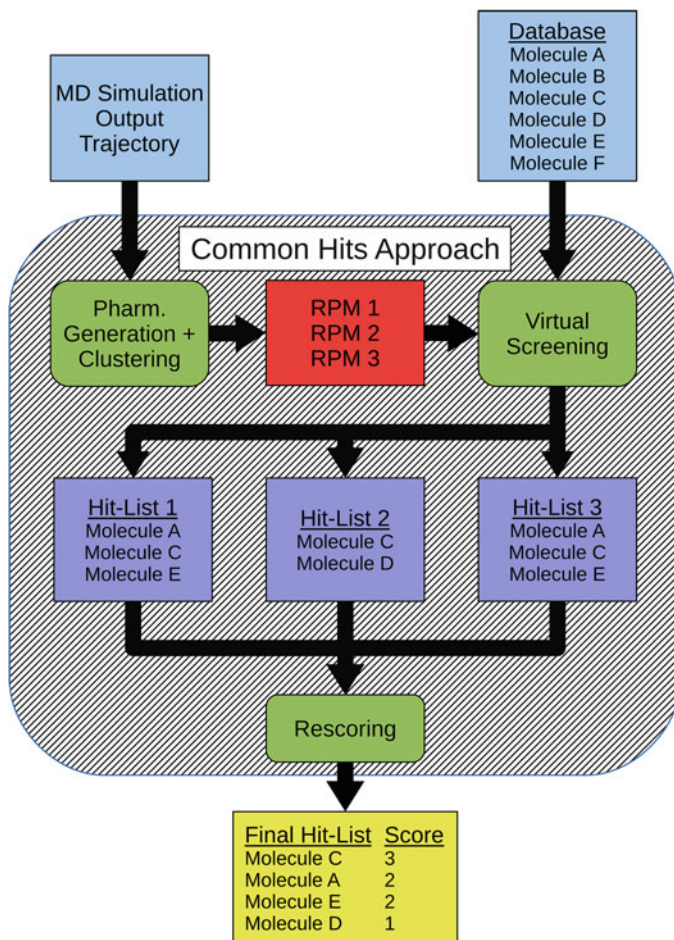


Fig. 12 Workflow of the “common hits approach” [100] starting with the MD output trajectory of a simulated ligand–receptor complex and a screening database. The final hit-list score of a molecule depends on the number of representative pharmacophore models (RPM) that matched the molecule

2.4.2 Pharmacophore-Based Interaction Fields

As briefly discussed in Sect. 2.2.2, molecular interaction field (MIF)-based approaches allow for an identification of ligand interaction hot spots in receptor binding sites. Once the preferred sites for particular types of nonbonded interactions are known they guide the placement of corresponding pharmacophoric features and thus allow generation of complete pharmacophore models [101, 102]. Especially when the three-dimensional structure of a ligand–receptor complex is unavailable and a direct perception of present ligand–target interactions thus not possible, employing a MIF-based method is the only way to obtain a reliable pharmacophore

hypothesis for the characterization of potential ligands. Molecular interaction fields moreover can be used for the guided optimization of existing ligands toward higher affinity by modifying ligand structures in a way that additional interaction sites pointed out by the MIF get captured [103]. A well-known program for the calculation of molecular interaction fields is GRID developed by Goodford et al. [35]. The GRID field evaluates a classical empirical energy function for a given probe molecule at each point of a regular grid that covers a region of interest of the target protein. Areas of the grid that have probe–target interaction energies above or below a certain user-defined threshold can then be visualized as isosurfaces using a suitable graphical display program. Areas with large negative energies indicate energetically favorable regions while those with large positive energies correspond to regions in which the probe molecule would encounter repelling forces. To obtain physically sound and accurate interaction energies, GRID uses an energy function containing terms that account for (1) hydrogen bonding (E_{hb}), (2) steric effects (E_{lj}), and (3) electrostatic attraction/repulsion (E_{el}). The total energy E is then calculated as the sum of the pairwise interaction energies between the atoms of the probe and the atoms of the target molecule:

$$E = \sum E_{\text{lj}} + \sum E_{\text{el}} + \sum E_{\text{hb}}$$

This energy function is essentially the same for every type of probe molecule. By calculating grid maps for multiple probes that are orthogonal in characteristics like charge distribution, size, and hydrogen bonding potential, it is possible to identify regions of the target receptor in which specific structural features of a ligand are more or less favorable. However, a drawback of this approach is that the calculated MIFs are specific for the molecular structure of the probe and the obtained interaction energies are always a mix of several contributions that cannot be separated easily. To be able to derive a set of favorable ligand features multiple grid calculations for different types of probes have to be performed and substantial postprocessing is required. Furthermore, calculations of physical interaction energies on the atomic level are computationally quite expensive and do not scale well with the size of the investigated system which makes them less suitable for an interaction analysis for whole MD trajectories.

If the primary goal is not the calculation of exact interaction energies but the identification of hot spots for certain types of ligand–target interactions, then the GRAIL approach [37] provides a viable alternative, which circumvents many of the problems mentioned before. In contrast to GRID, GRAIL works on a purely pharmacophoric representation of the target system and the probe is also just a single pharmacophoric feature of a particular type (representing, i.e., a hydrogen bond donor/acceptor, aromatic system, positive/negative ionizable group, hydrophobic group). The calculated interaction scores do not correspond to physical interaction energies as obtained by GRID, but rather reflect how well geometrical constraints like preferred distance and angle ranges are met when the probe feature gets placed at the grid points. The score at a particular grid point therefore gives an

insight as to how well a ligand feature of the probe type is able to interact with complementary features of the binding site. Performing grid calculations on a pharmacophoric feature level has two main advantages: (1) In comparison to atoms, due to the much smaller number of pharmacophoric features, the calculation of GRAIL grids is significantly faster than grid calculations on the atomic level. Therefore, dynamic information obtained by MD simulations can also be incorporated in the interaction hot-spot analysis, for even larger biological systems. (2) The number of probe and complementary target feature types is limited. Accordingly, the number of required grid calculations to capture all relevant nonbonded interactions for biological activity is also limited. Figure 13 shows examples for three different GRAIL grid maps covering the binding pocket region of CDK2. The bound inhibitor and the structure-based pharmacophore derived are shown to illustrate the good match of the identified interaction hot spots with the corresponding hydrophobic (yellow spheres), hydrogen bond acceptor (red arrows) and hydrogen bond donor (green arrows) features of the ligand pharmacophore.

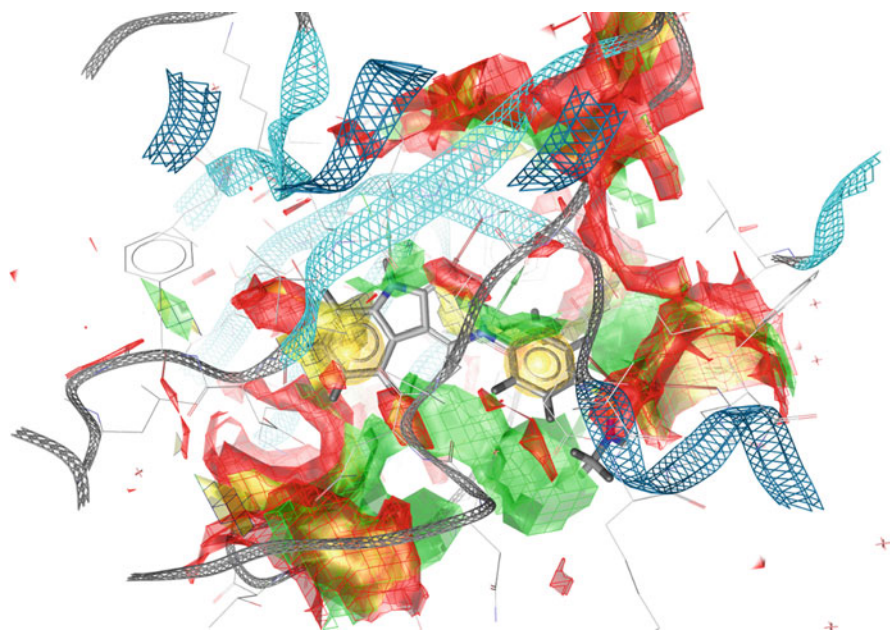


Fig. 13 Examples for GRAIL grid maps calculated for a CDK2/inhibitor complex (PDB-code: 1KE5). Red areas represent favorable ligand side regions for hydrogen bond acceptors, green areas for hydrogen bond donors, and yellow areas for lipophilic groups. The bound ligand and water molecules are ignored during the grid calculation and are only shown for reference

3 Application of Pharmacophores in Natural Product Research

Natural product-inspired design has become essential in designing tool compounds. There is growing interest in employing computational approaches for natural product-derived fragments. Chemotype-specific synthetic compound collections used to be the common approach in medicinal chemistry, in which compounds can be designed according to a specific target profile and aimed for a high degree of diversity [104, 105]. As natural products can serve as novel molecular scaffolds and therefore aid lead structure optimization, they have become an integral part of drug design, and can additionally stimulate ideas for synthesis [106]. Natural products do not always follow rules often applied in drug design (for instance, Lipinski's "Rule of Five" [107]), which opens the field to exploring powerful molecules in diverse chemical space. Recently some potential anticancer agents were described [108] possessing scaffolds inspired by natural products. These compounds are synthetically accessible, and the ability to link natural and synthetic compounds can be seen as a great benefit to medicinal chemistry [109]. Software has been made available to aim for natural product-derived chemical space, which correlate with bioactivity profiles [109, 110].

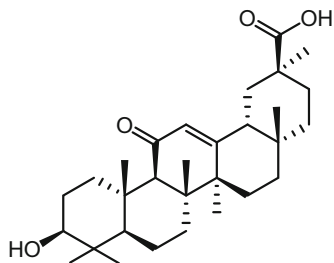
As this chapter has focused on three-dimensional pharmacophore modeling, some of the findings achieved through exploiting pharmacophore models for virtual screening purposes have been reviewed. By employing databases, curated from structures of natural products, the *in silico* approach can be successfully linked to the pharmacognostic profiles.

3.1 Screening for Selective Inhibitors of 11 β -Hydroxysteroid Dehydrogenase 1

One example of how pharmacophore modeling has successfully been employed to obtain potent inhibitors in natural product space was published in 2006 [111] and additional work on the same receptor was reported in 2010 [112].

11 β -Hydroxysteroid dehydrogenases (11 β -HSD) convert inactive 11-ketosteroids into the active, reduced 11 β derivatives. Inhibiting the enzyme 11 β -HSD1 has proved potentially important in the context of obesity, diabetes, and wound healing, as these conditions are glucocorticoid related. Enhanced expression of 11 β -HSD1 in adipose tissue of obese patients as well as in skeletal muscles of diabetic patients has been reported [113, 114].

Inhibition of closely related enzymes like 11 β -HSD2 or 17 β -HSD, however, leads to unwanted side effects, which is why the aim is to discover highly selective 11 β -HSD1 inhibitors. Pharmacophore modeling was employed to create a selective model to screen for potent inhibitors [115].

Fig. 14 Glycyrrhetic acid

The pharmacophore model utilized in this study was purely ligand based, since there was no X-ray structure available at the time the research was carried out. The ligand-based model could be retrieved due to known, active and selective 11β -HSD inhibitors [116, 117].

This pharmacophore model was used subsequently in order to screen databases to retrieve the aforementioned novel inhibitors. The steroidal core of the inhibitors was represented by hydrophobic and aromatic features in the pharmacophore model. Selectivity can be achieved by introducing substitutions to those core features that will be in charge of interacting with the environment by either forming hydrogen bonds, ionic interactions, or additional hydrophobic contacts.

Indeed, employing the ligand-based model and utilizing spatial restrictions derived from known inhibitors led to 15 novel hits with moderate to potent activity on 11β -HSD1.

In a second step, a nonselective model was built. However, a model for 11β -HSD2 proved difficult to establish the most potent and selective, known inhibitor of 11β -HSD2, which is a glycyrrhetic acid (Fig. 14) analog.

To incorporate information about 11β -HSD2 inhibitors, this was chosen as a first template for the nonselective model. Applying this model and its consecutive hits, the initial model could be refined and hits from the first hit list could be excluded, as they were not found to selectively target 11β -HSD1, but to be more promiscuous.

The final model was used for virtual screening of 12 commercially available compound collections to determine compounds suitable for biological testing. As many potent 11β -HSD inhibitors do not meet the requirements of orally available drugs, like, for instance, the Lipinski “Rule of Five” [107], these restrictions were not applied. In fact, the screen returned hits from both models and the test compounds proved to be selective for 11β -HSD1 over 11β -HSD2 for the 11β -HSD1 selective model and showed distinct inhibition of 11β -HSD2 for the nonselective model. Compounds with structural similarities to glycyrrhetic acid could be derived from the screen. Moreover, by using the 11β -HSD1 selective and nonselective models structurally diverse compounds could be obtained, which had not been published in an 11β -HSD1 inhibition context previously (Fig. 15).

These novel compounds are potential starting points for further optimization and possible therapeutic applications [111].

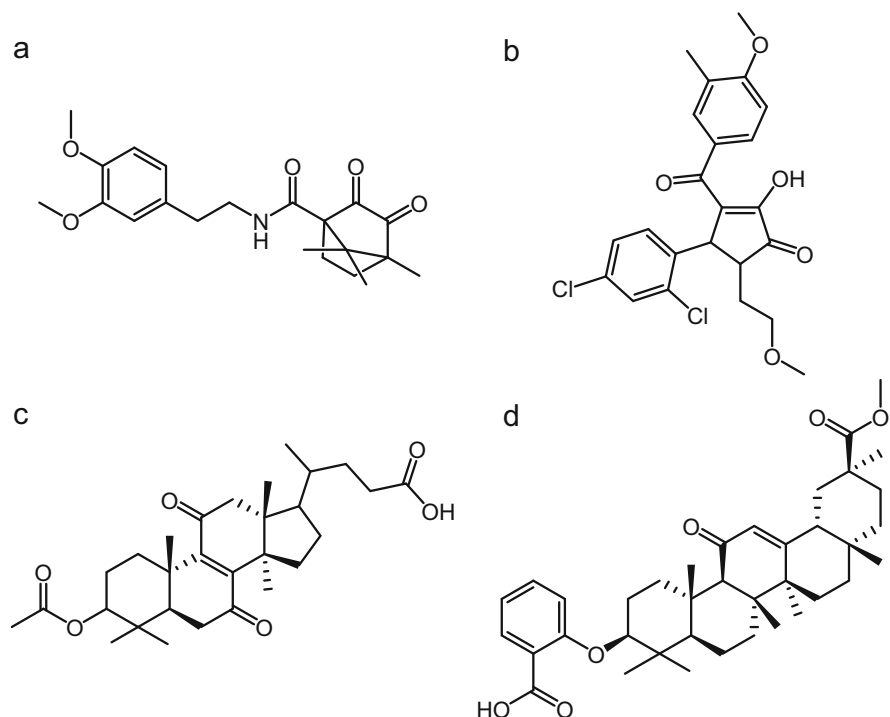


Fig. 15 Virtual screening hits (a) and (b): Structurally diverse compounds derived from the 11β -HSD1 selective model, with proved 11β -HSD1 selectivity in vitro. Compounds (c) and (d) showed significant inhibition of 11β -HSD2, in the range of $10\ \mu\text{M}$ or less. Compound (c) possesses a steroid-like structure and showed a 27-fold preference in inhibiting 11β -HSD1 over 11β -HSD2. Compound (d), however, is more GA like and inhibited 11β -HSD2 in vitro with a three- to fourfold preference

Following the work published in 2006 [111], Rollinger et al. further investigated the approach in 2010 [112] using the antidiabetic medicinal plant *Eriobotrya japonica* as a starting material for their studies. By employing the previously established pharmacophore model to screen the molecular database DIOS, which consists of 10,000 compounds from medicinal plants, 172 hits could be retrieved. A majority of the hits belonged to the scaffold class of the triterpenoids. One of the best scored hits in this context was the triterpene corosolic acid (Fig. 16e), which is a constituent of several herbal remedies. Furthermore, it is present in almond hulls [118], apple peel [119], and in the leaf extract of *E. japonica* [120, 121]. The latter is well known for its inhibitory effect on 11β -HSD1 and 11β -HSD2 [122]. Inhibitors of 11β -HSD1 have been reported to counteract the accumulation of visceral fat and reduce glucose blood level, and in addition, reduce metabolic risks in type-2 diabetes [123]. Further hits in the virtual screening were ursolic acid (Fig. 16b) as well as ursolic acid derivatives (Fig. 16c–d). After the screen, for further validation and examination of interactions, the output molecules were all docked into the 11β -

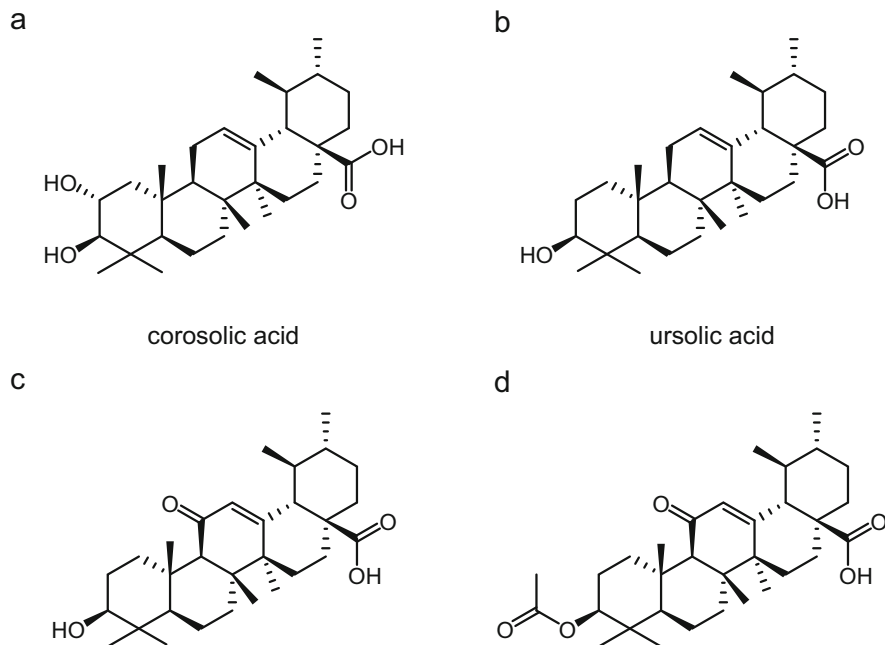


Fig. 16 Compound (a), corsolic acid, which was the first ranked compound in the hit list, showed an in vitro IC_{50} of $0.81 \mu M$, inhibiting 11β -HSD1 selectively. Compound (b), ursolic acid, with an IC_{50} of $1.9 \mu M$ was ranked just before the ursolic acid derivatives (c) and (d), displaying IC_{50} values of 2.06 and $1.35 \mu M$, respectively

HSD1 binding site. The docking algorithm allowed full ligand flexibility, partial protein flexibility, and water molecules in the binding domain were treated respectively.

After the placement, all chemical interactions were determined, utilizing the software LigandScout 3.0 [23]. The geometry, distance, and angles of the protein and the ligand in its vicinity were used to estimate the interactions. Subsequently, the structure–activity relationship could be assessed by interpreting chemical interactions, based on chemical functionalities.

The hits retrieved from the computational approach, exploiting pharmacophore modeling, exhibited selective inhibitors of 11β -HSD1 (Fig. 16). The hits displayed IC_{50} values between 0.8 and $2 \mu M$, which could be assessed in biological tests.

These results show how inhibitor-based pharmacophore modeling, in combination with virtual screening, can facilitate identifying novel, potent inhibitors in a natural product context.

3.2 Identification of Novel Natural Inhibitors of *Trypanosoma brucei* Glyceraldehyde-3-Phosphate Dehydrogenase

An example for structure-based pharmacophore modeling in conjugation with natural product databases was published in 2015 [124] and helped to identify natural inhibitors of *Trypanosoma brucei* glyceraldehyde-3-phosphate dehydrogenase (*Tb*GAPDH). *Trypanosoma brucei* is a protozoan parasite that can cause human African trypanosomiasis, also known as “sleeping sickness.” As a part of the “Neglected Tropical Diseases” classification by the World Health Organization, this infectious disease endangers more than 70 million sub-Saharan African people [125]. Suggested as potential drug target to deprive the parasite of energy supply [126], inhibitors of GAPDH represent promising trypanocidal agents [126–128]. Offering a vast structural diversity, natural products are known to exhibit high potential against protozoan infectious diseases [129, 130].

The database of natural product MEGx collection was selected to perform the virtual screening of the study. Among the 4803 available natural compounds, 700 were kept after several filtering steps such as the Lipinski’s Rule of Five [107] or limiting the number of stereocenters.

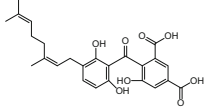
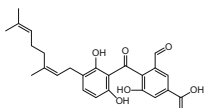
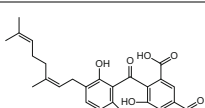
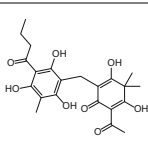

Three crystallographic structures (PDB-IDs: 2X0N, 3IDS, and 1GYP) of GAPDH from human pathogenic trypanosomatids were extracted from the Protein Data Bank and used to generate four structure-based pharmacophore models established on cocrystallized NAD⁺ and the last was built manually by analyzing an electrostatic map of the G-3-P site, due to the absence of the substrate.

Virtual screening was performed for each pharmacophore model. All resulting hits were docked in the respective binding site of the pharmacophore model from which they were retrieved. Based on their docking and virtual screening scores, 13 natural products were tested experimentally to validate the inhibition of GAPDH. The five compounds displaying enzyme inhibition superior to 50% at a concentration of 50 μ M were selected and their IC_{50} values were determined.

Three compounds were geranylated benzophenone derivatives extracted from the fungus *Geniculosporium* sp., one compound was a flavaspodic acid analog extracted from the fern *Dryopteris crassirhizoma*, and the last was a tetradecane derivative extracted from the tree *Grevillea whiteana*. Structures of the natural compounds are depicted in Table 3. All five tested compounds exhibited inhibition of GAPDH, displaying IC_{50} values below 30 μ M, while two of the geranylated benzophenone derivatives exhibited an IC_{50} value lower than 8 μ M.

Considering the significant rate of experimentally confirmed inhibitors from more than 4800 initial natural products, structure-based pharmacophore modeling thus helped to selectively identify five new *Tb*GAPDH inhibitors.

Table 3 List of Five Natural Products Exhibiting More Than 50% *Tb*GAPDH Inhibition at a Concentration of 50 μ M

Compound structure	% of <i>Tb</i> GAPDH inhibition at $c = 50 \mu$ M	IC_{50} (μ M)	Compound class
	66	24.56 ± 1.03	Geranylated benzophenone derivatives
	98	4.73 ± 1.03	
	>90	6.68 ± 1.04	
	88	21.97 ± 1.03	Flavaspidic acid AB
	>90	22.79 ± 1.01	Tetradecane derivative

References

1. Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA (1998) Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *J Macromol Sci A Pure Appl Chem* 70:1129
2. Wermuth CG (2006) Pharmacophores: historical perspective and viewpoint from a medicinal chemist. In: Langer T, Hoffmann RD (eds) *Methods and principles in medicinal chemistry*, vol 32. Wiley-VCH, Weinheim, p 1
3. Langley JN (1878) On the physiology of the salivary secretion. *J Physiol* 1:339
4. Langley JN (1905) On the reaction of cells and of nerve-endings to certain poisons, chiefly as regards the reaction of striated muscle to nicotine and to curari. *J Physiol* 33:374
5. Ehrlich P, Morgenroth J (1900) Über Hämolyse. *Dritte Mittheilung Berl Klin Wschr* 37:453
6. Maehle A-H, Prüll C-R, Halliwell RF (2002) The emergence of the drug receptor theory. *Nat Rev Drug Discov* 1:637
7. Albert A (1985) *Selective toxicity: the physico-chemical basis of therapy*. Springer, Netherlands
8. Fischer E (1894) Einfluss der Configuration auf die Wirkung der Enzyme. *Ber Dtsch Chem Ges* 27:2985

9. Woods DD (1940) The relation of *p*-aminobenzoic acid to the mechanism of the action of sulphanilamide. *Br J Exp Pathol* 21:74
10. Woods DD, Fildes P (1940) The anti-sulphanilamide activity (in vitro) of *p*-aminobenzoic acid and related compounds. *Chem Ind* 59:133
11. Dodds EC, Lawson W (1938) Molecular structure in relation to oestrogenic activity. Compounds without a phenanthrene nucleus. *Proc R Soc Lond B Biol Sci* 125:222
12. Easson LH, Stedman E (1933) Studies on the relationship between chemical constitution and physiological action: molecular dissymmetry and physiological activity. *Biochem J* 27:1257
13. Gund P (2000) Evolution of the pharmacophore concept in pharmaceutical research. In: Güner OF (ed) *Pharmacophore perception, development, and use in drug design*. International University Line, La Jolla, CA, p 5
14. Matthews DA, Alden RA, Bolin JT, Freer ST, Hamlin R, Hol WGJ, Kisliuk RL, Pastore EJ, Plante LT, Xuong N Kraut J (1978) Dihydrofolate reductase from *Lactobacillus casei*. X-Ray structure of the enzyme methotrexate complex. *J Biol Chem* 253:6946
15. Wolfenden R (1976) Transition state analog inhibitors and enzyme catalysis. *Annu Rev Biophys Bioeng* 5:271
16. Gund P (1979) Pharmacophoric pattern searching and receptor mapping. In: Hess H-J (ed) *Annual reports in medicinal chemistry*, vol 14. Academic, New York, p 299
17. Humblet C, Marshall GR (1980) Pharmacophore identification and receptor mapping. In: Hess H-J (ed) *Annual reports in medicinal chemistry*, vol 15. Academic, New York, p 267
18. Marshall GR, Barry CD, Bosshard HE, Dammkoehler RA, Dunn DA (1979) In: Olson EC, Christoffersen RE (eds) *The conformational parameter in drug design: The active analog approach*, vol 112. American Chemical Society Books, Washington, DC, p 205
19. Greene J, Kahn S, Savoj H et al (1994) Chemical function queries for 3D database search. *J Chem Inf Comput Sci* 34:1297
20. Discovery Studio Predictive Science Application | Dassault Systèmes BIOVIA. <https://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/>. Accessed 7 Feb 2019
21. Phase | Schrödinger. <https://www.schrodinger.com/phase>. Accessed 7 Feb 2019
22. Molecular operating environment (MOE) | CCG Inc. https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm. Accessed 7 Feb 2019
23. LigandScout | IntelLigand GmbH. <http://www.inteligand.com/ligandscout/>. Accessed 7 Feb 2019
24. Wolber G, Kosara R (2006) Pharmacophores from macromolecular complexes with LigandScout. In: Langer T, Hoffmann RD (eds) *Methods and principles in medicinal chemistry*, vol 32. Wiley-VCH, Weinheim, p 13
25. Wolber G, Langer T (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model* 45:160
26. Wolber G, Seidel T, Bendix F, Langer T (2008) Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov Today* 13:23
27. Krovat EM, Langer T (2003) Non-peptide angiotensin II receptor antagonists: chemical feature based pharmacophore identification. *J Med Chem* 46:716
28. Williams MA, Ladbury JE (2008) Hydrogen bonds in protein-ligand complexes. *Protein Science Encyclopedia*: online 137
29. Böhm H-J, Klebe G, Kubinyi H (1996) Protein-ligand-Wechselwirkungen. In: Böhm HJ, Klebe G, Kubinyi H (eds) *Wirkstoffdesign*. Spektrum Akademischer, Heidelberg, p 95
30. Ma JC, Dougherty DA (1997) The cation- π interaction. *Chem Rev* 97:1303
31. Waters ML (2002) Aromatic interactions in model systems. *Curr Opin Chem Biol* 6:736
32. Böhm H-J, Klebe G, Kubinyi H (1996) Metalloprotease-Hemmer. In: Böhm HJ, Klebe G, Kubinyi H (eds) *Wirkstoffdesign*. Spektrum. Akademischer Verlag, Heidelberg, p 505
33. Englert L, Silber K, Steuber H, Brass S, Over B, Gerber HD, Heine A, Diederich WE, Klebe G (2010) Fragment-based lead discovery: screening and optimizing fragments for thermolysin inhibition. *ChemMedChem* 5:930

34. Leach AR, Gillet VJ, Lewis RA, Taylor R (2010) Three-dimensional pharmacophore methods in drug discovery. *J Med Chem* 53:539
35. Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 28:849
36. Wade RC, Goodford PJ (1993) Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *J Med Chem* 36:148
37. Schuetz DA, Seidel T, Garon A, Martini R, Körbel M, Ecker GF, Langer T (2018) GRAIL: GRIDs of phARmacophore Interaction fieLds. *J Chem Theory Comput* 14:4958
38. Böhm H-J (1992) The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J Comput Aided Mol Des* 6:61
39. Verdonk ML, Cole JC, Taylor R (1999) SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *J Mol Biol* 289:1093
40. Kirchhoff PD, Brown R, Kahn S, Waldman M, Venkatachalam CM (2001) Application of structure-based focusing to the estrogen receptor. *J Comput Chem* 22:993
41. Venkatachalam CM, Kirchhoff P, Waldman M (2000) Receptor-based pharmacophore perception and modeling. In: Güner OF (ed) *Pharmacophore perception, development, and use in drug design*. International University Line, La Jolla, CA, p 341
42. Dixon SL (2010) Pharmacophore methods. In: Merz KM Jr, Ringe D, Reynolds CH (eds) *Drug design: structure- and ligand-based approaches*. Cambridge University Press, Cambridge, p 137
43. Poptodorov K, Luu T, Hoffmann RD (2006) Pharmacophore model generation software tools. In: Langer T, Hoffmann RD (eds) *Methods and principles in medicinal chemistry*, vol 32. Wiley-VCH, Weinheim, p 17
44. Triballeau N, Bertrand H-O, Achner F (2006) Are you sure you have a good model? In: Langer T, Hoffmann RD (eds) *Methods and principles in medicinal chemistry*, vol 32. Wiley-VCH, Weinheim, p 325
45. Martin YC (1992) 3D database searching in drug design. *J Med Chem* 35:2145
46. Manallack DT (1996) Getting that hit: 3D database searching in drug discovery. *Drug Discov Today* 1:231
47. Clark DE, Westhead DR, Sykes RA, Murray CW (1996) Active-site-directed 3D database searching: pharmacophore extraction and validation of hits. *J Comput Aided Mol Des* 10:397
48. Good AC, Mason JS (1996) Three-dimensional structure database searches. *Rev Comput Chem* 67
49. Hurst T (1994) Flexible 3D searching: The directed tweak technique. *J Chem Inf Comput Sci* 34:190
50. Wolber G, Dornhofer AA, Langer T (2006) Efficient overlay of small organic molecules using 3D pharmacophores. *J Comput Aided Mol Des* 20:773
51. Laggner C, Wolber G, Kirchmair J, Schuster D, Langer T (2008) Pharmacophore-based virtual screening in drug discovery. In: *Chemoinformatics approaches to virtual screening*. The Royal Society of Chemistry, London, p 76
52. Sheridan RP, Kearsley SK (2002) Why do we need so many chemical similarity search methods? *Drug Discov Today* 7:903
53. Johnson MA, Maggiora GM (1990) *Concepts and applications of molecular similarity*. Wiley, New York
54. Leach AR (2001) *Molecular modelling: principles and applications*. Pearson Education, London
55. Zhu F, Agrafiotis DK (2007) Recursive distance partitioning algorithm for common pharmacophore identification. *J Chem Inf Model* 47:1619
56. Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Freisner RA (2006) PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J Comput Aided Mol Des* 20:647

57. Wolber G, Seidel T, Bendix F (2010) 3D pharmacophore alignments: does improved geometric accuracy affect virtual screening performance? *J Cheminform* 2:O10
58. Evers A, Hessler G, Matter H, Klabunde T (2005) Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols. *J Med Chem* 48:5448
59. Brint AT, Willett P (1987) Algorithms for the identification of three-dimensional maximal common substructures. *J Chem Inf Comput Sci* 27:152
60. Barnum D, Greene J, Smellie A, Sprague P (1996) Identification of common functional configurations among molecules. *J Chem Inf Comput Sci* 36:5631
61. Lemmen C, Lengauer T (2000) Computational methods for the structural alignment of molecules. *J Comput Aided Mol Des* 14:215
62. Seidel T, Bendix F, Wolber G (2010) Strategies for 3D pharmacophore-based virtual screening. *Drug Discov Today Technol* 7:e203–e270
63. Güner OF (2000) Pharmacophore perception. Development and use in drug design. International University Line, La Jolla, CA
64. Kirchmair J, Ristic S, Eder K, Markt P, Wolber G, Laggner C, Langer T (2007) Fast and efficient in silico 3D screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J Chem Inf Model* 47:2182
65. Langer T, Wolber G (2004) Pharmacophore definition and 3D searches. *Drug Discov Today Technol* 1:203
66. Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, New York
67. Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H-O (2005) Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J Med Chem* 48:2534
68. Schneider G, Fechner U (2005) Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov* 4:649
69. Gillet VJ, Johnson AP, Mata P, Sike S (1990) Automated structure design in 3D. *Tetrahedron Comput Methodol* 3:681
70. Nishibata Y, Itai A (1991) Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron* 47:8985
71. Tschinke V, Cohen NC (1993) The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypotheses. *J Med Chem* 36:3863
72. Pearlman DA, Murcko MA (1993) CONCEPTS: new dynamic algorithm for de novo drug suggestion. *J Comput Chem* 14:1184
73. Gillet V, Johnson AP, Mata P, Sike S, Williams P (1993) SPROUT: a program for structure generation. *J Comput Aided Mol Des* 7:127
74. Eisen MB, Wiley DC, Karplus M, Hubbard RE (1994) HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins* 19:199
75. DeWitte RS, Shakhnovich EI (1996) SMOG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J Am Chem Soc* 118:11733
76. Pearlman DA, Murcko MA (1996) CONCERTS: dynamic connection of fragments as an approach to de novo ligand design. *J Med Chem* 39:1651
77. Douguet D, Thoreau E, Grassy G (2000) A genetic algorithm for the automated generation of small organic molecules: drug design using an evolutionary algorithm. *J Comput Aided Mol Des* 14:449
78. Wang R, Gao Y, Lai L (2000) LigBuilder: a multi-purpose program for structure-based drug design. *Mol Mod Ann* 6:498
79. Schneider G, Lee ML, Stahl M, Schneider P (2000) De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J Comput Aided Mol Des* 14:487

80. Zhu J, Fan H, Liu H, Shi Y (2001) Structure-based ligand design for flexible proteins: application of new F-DycoBlock. *J Comput Aided Mol Des* 15:979
81. Pegg SC, Haresco JJ, Kuntz ID (2001) A genetic algorithm for structure-based de novo design. *J Comput Aided Mol Des* 15:911
82. Vinkers HM, de Jonge MR, Daeyaert FFD, Daeyaert FF, Heeres J, Koymans LM, van Lenthe JH, Lewi PJ, Timmerman H, Van Aken K, Janssen PA (2003) SYNOPSIS: SYNthesize and OPTimize System in Silico. *J Med Chem* 46:2765
83. Brown N, McKay B, Gilardoni F, Gasteiger J (2004) A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *ChemInform* 35:1079
84. Pierce AC, Rao G, Bemis GW (2004) BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease. *J Med Chem* 47:2768
85. Huang Q, Li L-L, Yang S-Y (2010) PhDD: a new pharmacophore-based de novo design method of drug-like molecules combined with assessment of synthetic accessibility. *J Mol Graph Model* 28:775
86. Mirjalili V, Noyes K, Feig M (2014) Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. *Proteins* 82(Suppl 2):196
87. Whitesides GM, Krishnamurthy VM (2005) Designing ligands to bind proteins. *Q Rev Biophys* 38:385
88. Yang S-Y (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov Today* 15:444
89. Wu F, Xu T, He G, Ouyang L, Peng C, Song Z, Xiang M (2012) Discovery of novel focal adhesion kinase inhibitors using a hybrid protocol of virtual screening approach based on multicomplex-based pharmacophore and molecular docking. *Int J Mol Sci* 13:15668
90. Zou J, Xie H-Z, Yang S-Y, Chen JJ, Ren JX, Wei YQ (2008) Towards more accurate pharmacophore modeling: multicomplex-based comprehensive pharmacophore map and most-frequent-feature pharmacophore model of CDK2. *J Mol Graph Model* 27:430
91. Choudhury C, Priyakumar UD, Sastry GN (2015) Dynamics based pharmacophore models for screening potential inhibitors of mycobacterial cyclopropane synthase. *J Chem Inf Model* 55:848
92. Sohn Y-S, Park C, Lee Y, Thangapandian S, Kim Y, Kim HH, Suh JK, Lee KW (2013) Multi-conformational dynamic pharmacophore modeling of the peroxisome proliferator-activated receptor γ for the discovery of novel agonists. *J Mol Graph Model* 46:1
93. Thangapandian S, John S, Arooj M, Lee KW (2012) Molecular dynamics simulation study and hybrid pharmacophore model development in human LTA4H inhibitor design. *PLoS One* 7:e34593
94. Thangapandian S, John S, Lee Y, Kim S, Lee KW (2011) Dynamic structure-based pharmacophore model development: a new and effective addition in the histone deacetylase 8 (HDAC8) inhibitor discovery. *Int J Mol Sci* 12:9440
95. Spyrikis F, Benedetti P, Decherchi S, Rocchia W, Cavalli A, Alcaro S, Ortuso F, Baroni M, Cruciani G (2015) A pipeline to enhance ligand virtual screening: integrating molecular dynamics and fingerprints for ligand and proteins. *J Chem Inf Model* 55:2256
96. Sydow D (2015) Dynophores: novel dynamic pharmacophores. Humboldt-Universität zu Berlin, Lebenswissenschaftliche Fakultät
97. Sinko W, Lindert S, McCammon JA (2013) Accounting for receptor flexibility and enhanced sampling methods in computer-aided drug design. *Chem Biol Drug Des* 81:41
98. Plattner N, Noé F (2015) Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat Commun* 6:7653
99. Wieder M, Perricone U, Boresch S, Seidel T, Langer T (2016) Evaluating the stability of pharmacophore features using molecular dynamics simulations. *Biochem Biophys Res Commun* 470:685
100. Wieder M, Garon A, Perricone U, Boresch S, Seidel T, Almerico AM, Langer T (2017) Common hits approach: combining pharmacophore modeling and molecular dynamics simulations. *J Chem Inf Model* 57:365

101. Ortuso F, Langer T, Alcaro S (2006) GBPM: GRID-based pharmacophore model: concept and application studies to protein–protein recognition. *Bioinformatics* 22:1449
102. Mortier J, Dhakal P, Volkamer A (2018) Truly target-focused pharmacophore modeling: a novel tool for mapping intermolecular surfaces. *Molecules* 23:99E1959
103. Kastenholz MA, Pastor M, Cruciani G, Haaksma EEJ, T I F (2000) GRID/CPCA: a new computational tool to design selective ligands. *J Med Chem* 43:3033
104. Filer CN (2008) Book review of molecular design. Concepts and applications molecular design. Concepts and applications. By Schneider G, Baringhaus K-H. *J Med Chem* 51:7020
105. Schreiber SL (2000) Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* 287:1964
106. Grabowski K, Baringhaus K-H, Schneider G (2008) Scaffold diversity of natural products: inspiration for combinatorial library design. *Nat Prod Rep* 25:892
107. Lipinski CA (2004) Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol* 1:337
108. Elumalai N, Berg A, Natarajan K, Scharow A, Berg T (2015) Nanomolar inhibitors of the transcription factor STAT5b with high selectivity over STAT5a. *Angew Chem Int Ed Engl* 54:4758
109. Rodrigues T, Reker D, Schneider P, Schneider G (2016) ChemInform abstract: counting on natural products for drug design. *ChemInform* 47. <https://doi.org/10.1002/chin.201630259>
110. Larsson J, Gottfries J, Muresan S, Backlund A (2007) ChemGPS-NP: tuned for navigation in biologically relevant chemical space. *J Nat Prod* 70:789
111. Schuster D, Maurer EM, Laggner C, Nashev LG, Wilckens T, Langer T, Odermatt A (2006) The discovery of new 11 β -hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. *J Med Chem* 49:3454
112. Rollinger JM, Kratschmar DV, Schuster D, Pfisterer PH, Gumy C, Aubry EM, Brandstötter S, Stuppner H, Wolber G, Odermatt A (2010) 11 β -Hydroxysteroid dehydrogenase 1 inhibiting constituents from *Eriobotrya japonica* revealed by bioactivity-guided isolation and computational approaches. *Bioorg Med Chem* 18:1507
113. Abdallah BM, Beck-Nielsen H, Gaster M (2005) Increased expression of 11 β -hydroxysteroid dehydrogenase type 1 in type 2 diabetic myotubes. *Eur J Clin Invest* 35:627
114. Kannisto K, Pietiläinen KH, Ehrenborg E, Rissanen A, Kaprio J, Hamsten A, H I Y-J (2004) Overexpression of 11 β -hydroxysteroid dehydrogenase-1 in adipose tissue is associated with acquired obesity and features of insulin resistance: studies in young adult monozygotic twins. *J Clin Endocrinol Metab* 89:4414
115. Schuster D, Wolber G (2010) Identification of bioactive natural products by pharmacophore-based virtual screening. *Curr Pharm Des* 16:1666
116. Barf T, Vallgård A, Emond R, Häggström C, Kurz G, Nygren A, Larwood V, Mosialou E, Axelsson K, Olsson R, Engblom L, Edling N, Rönquist-Nii Y, Ohman B, Alberts P, Abrahamsén L (2002) Arylsulfonamidothiazoles as a new class of potential antidiabetic drugs. Discovery of potent and selective inhibitors of the 11 β -hydroxysteroid dehydrogenase type 1. *J Med Chem* 45:3813
117. Vicker N, Su X, Ganeshpillai D, Smith A, Purohit A, Reed MJ, Potter BVL (2007) Novel non-steroidal inhibitors of human 11 β -hydroxysteroid dehydrogenase type 1. *J Steroid Biochem Mol Biol* 104:123
118. Amico V, Barresi V, Condorelli D, Spatafora C, Tringali C (2006) Antiproliferative terpenoids from almond hulls (*Prunus dulcis*): identification and structure-activity relationships. *J Agric Food Chem* 54:810
119. He X, Liu RH (2007) Triterpenoids isolated from apple peels have potent antiproliferative activity and may be partially responsible for apple's anticancer activity. *J Agric Food Chem* 55:4366
120. Gilar M (2001) Analysis and purification of synthetic oligonucleotides by reversed-phase high-performance liquid chromatography with photodiode array and mass spectrometry detection. *Anal Biochem* 298:196

121. Liang ZZ, Aquino R, de Feo V, De Simone F, Pizza C (1990) Polyhydroxylated triterpenes from *Eriobotrya japonica*. *Planta Med* 56:330
122. Gummy C, Thurnbichler C, Aubry EM, Balazs Z, Pfisterer P, Baumgartner L, Stuppner H, Odermatt A, Rollinger JM (2009) Inhibition of 11 β -hydroxysteroid dehydrogenase type 1 by plant extracts used as traditional antidiabetic medicines. *Fitoterapia* 80:200
123. Hughes KA, Webster SP, Walker BR (2008) 11- β -Hydroxysteroid dehydrogenase type 1 (11 β -HSD1) inhibitors in Type 2 diabetes mellitus and obesity. *Exp Opin Invest Drugs* 17:481
124. Herrmann F, Lenz M, Jose J, Kaiser M, Brun R, Schmidt TJ (2015) In silico identification and in vitro activity of novel natural inhibitors of *Trypanosoma brucei* glyceraldehyde-3-phosphate-dehydrogenase. *Molecules* 20:16154
125. World Health Organization (2015) Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected tropical diseases 2015. World Health Organization
126. Gualdrón-López M, Michels PAM, Quiñones W, Cáceres AJ, Avilán L, Concepción JL (2013) Function of glycosomes in the metabolism of trypanosomatid parasites and the promise of glycosomal proteins as drug targets. In: Jäger T, Koch O, Fiohé L (eds) *Trypanosomatid diseases: molecular routes to drug discovery*. Wiley-VCH, Weinheim, p 121
127. Cáceres AJ, Michels PAM, Hannaert V (2010) Genetic validation of aldolase and glyceraldehyde-3-phosphate dehydrogenase as drug targets in *Trypanosoma brucei*. *Mol Biochem Parasitol* 169:50
128. Schuster R, Holzhutter H-G (1995) Use of mathematical models for predicting the metabolic effect of large-scale enzyme activity alterations. Application to enzyme deficiencies of red blood cells. *Eur J Biochem* 229:403
129. Schmidt TJ, Khalid SA, Romanha AJ, Alves TM, Biavatti MW, Brun R, Da Costa FB, de Castro SL, Ferreira VF, de Lacerda MV, Lago JH, Leon LL, Lopes NP, das Neves Amorim RC, Niehues M, Ogungbe IV, Pohlit AM, Scotti MT, Setzer WN, de N C Soeiro M, Steindel M, Tempone AG (2012) The potential of secondary metabolites from plants as drugs or leads against protozoan neglected diseases – Part I. *Curr Med Chem* 19:2128
130. Schmidt TJ, Khalid SA, Romanha AJ, Alves TM, Biavatti MW, Brun R, Da Costa FB, de Castro SL, Ferreira VF, de Lacerda MV, Lago JH, Leon LL, Lopes NP, das Neves Amorim RC, Niehues M, Ogungbe IV, Pohlit AM, Scotti MT, Setzer WN, de N C Soeiro M, Steindel M, Tempone AG (2012) The potential of secondary metabolites from plants as drugs or leads against protozoan neglected diseases – Part II. *Curr Med Chem* 19:2176



Thomas Seidel studied chemistry at the Vienna University of Technology where he received a Ph.D. in synthetic organic chemistry in 2003. From 2004 until 2005 he worked as a postdoctoral research assistant in the group of Prof. Johann Gasteiger at the Computer Chemistry Center/University of Erlangen-Nuremberg. The main focus of his research was the development of methods and algorithms to assess the synthetic accessibility of compounds generated by de novo design. After his postdoctoral period he returned to Vienna for a position at the company IntelLigand GmbH where he developed cheminformatics software for pharmacophore modeling and virtual screening. Since 2014, Dr. Seidel has been a senior postdoc at the Department of Pharmaceutical Chemistry/University of Vienna where he leads the cheminformatics research group. The main topics of his research are the development and implementation of novel algorithms for pharmacophore-based drug development, the advancement of the pharmacophore concept, and the application of molecular dynamics simulations and machine-learning methods in the context of molecular design.



Doris Alexandra Schuetz studied pharmacy at the University of Vienna and performed her Master's project in Melbourne, Australia, where she worked in a R&D laboratory for Boron Molecular as a synthetic chemist. She continued her research in the Pharmacoinformatics group at the University of Vienna, and received her Ph.D. in 2018. During her doctorate, research was mainly focused on kinetics in drug discovery. One of her main interests is structure-based drug design, involving molecular dynamics simulations. She spent time at the Department of Pharmacy and Biotechnology in Bologna to specialize in Molecular Dynamics simulations for kinetic studies in 2016. After receiving her Ph.D., Dr. Schuetz joined Inte:Ligand GmbH as an application scientist for project work. She has worked on the development of a pharmacophore-based interaction field method for the prediction of interaction hot spots in apo binding sites. In September 2018, Dr. Schuetz joined the computational chemistry team at the Institute for Research in Immunology and Cancer (IRIC) in Montréal, Canada. At IRIC, based in the Université de Montréal, she is working as a postdoctoral researcher, with a strong focus on structure-based drug design. She is mainly working on protein–protein interfaces, aiming to integrate structural information into small-molecule and macrocycle design in a cancer context.



Arthur Garon studied chemistry and biology at the University of Clermont-Ferrand and obtained an M.Sc. degree in cheminformatics at the University of Strasbourg in 2016. He has worked in the Cheminformatics Research Group of the Pharmaceutical Chemistry Department at the University of Vienna, where he started his Ph.D. in 2017 in collaboration with the University of Vienna, InteLigand GmbH, and Laboratoires Servier. The research being carried out during his doctorate involves the development and implementation of new methods for pharmacophore-based drug development. He is working on the extraction and analysis of pharmacophore-related information derived from crystallographic structures and molecular dynamics simulations.



Thierry Langer holds an M.Sc. degree in Pharmacy (1988) and a Ph.D. in Pharmaceutical Chemistry (1991) from the University of Vienna. He began his academic career at Leopold-Franzens-University of Innsbruck in 1992 after completing a postdoctoral fellowship at the Université Louis Pasteur, Strasbourg, France with Prof. C.-G. Wermuth. In 1997, he was appointed Associate Professor of Pharmaceutical Chemistry at the University of Innsbruck. In 2003, with colleagues, he founded the company Inte:Ligand GmbH, which develops scientific software for computer-aided molecular design, and served as the CEO until 2008. Then he was appointed CEO of Prestwick Chemical, Inc., a world-renowned contract research organization that specializes in medicinal chemistry services, located in Strasbourg, France. Under his leadership, several drug discovery programs in different research target sectors successfully progressed into preclinical and clinical development. In 2013, he was nominated as full professor for medicinal chemistry at the University of Vienna, where he currently heads the Department of Pharmaceutical Chemistry at the Faculty of Life Sciences.

Cheminformatic Analysis of Natural Product Fragments



Daniel Reker

Contents

1	Introduction	143
2	Sources of Natural Product Fragments	145
2.1	Filtering	146
2.2	Virtual Fragmentation	147
2.3	Scaffolds	148
3	Properties of Natural Product Fragments	149
3.1	Chemical and Physical Properties	150
3.2	Spatial Properties	150
3.3	Natural Fragments from Different Sources	151
3.4	Commercial Availability and Synthetic Tractability	152
3.5	Problematic Natural Fragments	154
4	Applications of Natural Product Fragments	156
4.1	Predicting Biomacromolecular Targets of Natural Fragments	156
4.2	Collection Design	160
4.3	Analysis of Natural Product-Likeness	165
5	Concluding Remarks and Outlook	166
	References	167

1 Introduction

Natural products have been fueling drug discovery pipelines for decades [1, 2]. However, many challenging hurdles have hampered the straightforward application of complex natural product structures for drug discovery, such as a lack of synthetic accessibility for large-scale production [3, 4] as well as their unknown and insufficiently predictable polypharmacological properties [5, 6]. Conversely, natural

D. Reker (✉)

Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology,
Cambridge, MA, USA

e-mail: reker@mit.edu

© Springer Nature Switzerland AG 2019

A. D. Kinghorn, H. Falk, S. Gibbons, J. Kobayashi, Y. Asakawa, J.-K. Liu (eds.),

Progress in the Chemistry of Organic Natural Products, Vol. 110,

https://doi.org/10.1007/978-3-030-14632-0_5

143

product fragments overcome many of such shortcomings and provide privileged structures with ample opportunities for optimization and derivatization into synthetically accessible mimetics with known or predictable biological effects [4, 5, 7, 8]. It has been shown that natural product fragments can provide key substructures to bias a compound collection toward biological activity [9, 10], and it has been argued that fragment-sized natural products and natural fragments constitute some of the most relevant of natural products for drug discovery and development [11]. The biosynthesis of a large natural product often relies on the synthesis of fragment-like building blocks, such that these structures might present biologically motivated handles to further explore them as chemical probes and drug leads [12].

Furthermore, natural fragments provide innovative as well as structurally and spatially intricate molecular probes for fragment-based drug discovery [4, 8]. This is particularly relevant given the much smaller size of chemical fragment space: since there are orders of magnitude less possible fragment structures compared to the unfathomable larger number of possible organic molecules [10], a smaller fragment collection might be capable of spanning the complete chemical (fragment) space with sufficient resolution [13]. This implies that the high-throughput testing of fragments and their medicinal chemistry is more likely to enable the design of bioactive compounds with optimal activity profile [14]. Indeed, focusing on natural fragments benefits from the advantages of chemically diverse and complex natural product structures, while simultaneously harnessing fragment-based drug discovery for finding smaller structures that fit optimally into a binding pocket and can then be rationally further optimized [13]. The potential for such optimization is additionally attested to by orthogonal work on privileged fragments and scaffolds [5, 15], which are notorious for binding certain target classes and families. This advocates for the selection of molecular substructures that contribute most significantly to the desired activity of biologically active molecules—and natural product fragments constitute a prime resource to study these [16].

With these vital benefits of natural fragments in mind, it is important to realize that cheminformatic and bioinformatic approaches have been instrumental in generating and curating databases of natural product-derived fragments [4, 8, 17], analyzing and stratifying their properties [17, 18], as well as guiding their applications for drug discovery and chemical biology [5, 16, 19–23]. Many *in silico* tools exist that have been validated with impressive success in the context of natural fragments to support their generation, property and polypharmacology predictions, and derivatization, among others [9, 15, 24, 25]. This contribution discusses various challenges and opportunities occurring at different stages of computer-guided natural product fragment research, currently available computational tools to address these, as well as outstanding research questions and the prospective impact of computational workflows on natural product-based drug discovery and chemical biology.

2 Sources of Natural Product Fragments

To utilize natural product fragments in drug discovery and pharmaceutical research, it is of utmost importance to source a large set of reliable and meaningful chemical structures for further analysis or screening. A common selection criterion for natural products and fragments is compound availability [4], and multiple chemical vendors now supply chemical structures of natural product fragments that they offer commercially for academic and industrial researchers to fuel natural fragment-based screening efforts [26–29]. However, the few currently available collections are limited in the number of included structures and are likely biased toward more readily accessible fragments or have undergone other external filtering criteria such as drug- or lead-likeness [30–32]. This can unknowingly distort the compound collection and thereby dramatically impact the trajectory of a given project. For cheminformatic analysis or downstream drug and chemical probe discovery, the automated extraction of fragments from vast natural product collections is a fruitful strategy to generate large datasets of natural product fragments. Multiple orthogonal strategies exist (Fig. 1) that enable rapid, chemically meaningful, and reproducible generation of natural fragment collections [4, 7–9, 17]. The strategy of choice will be

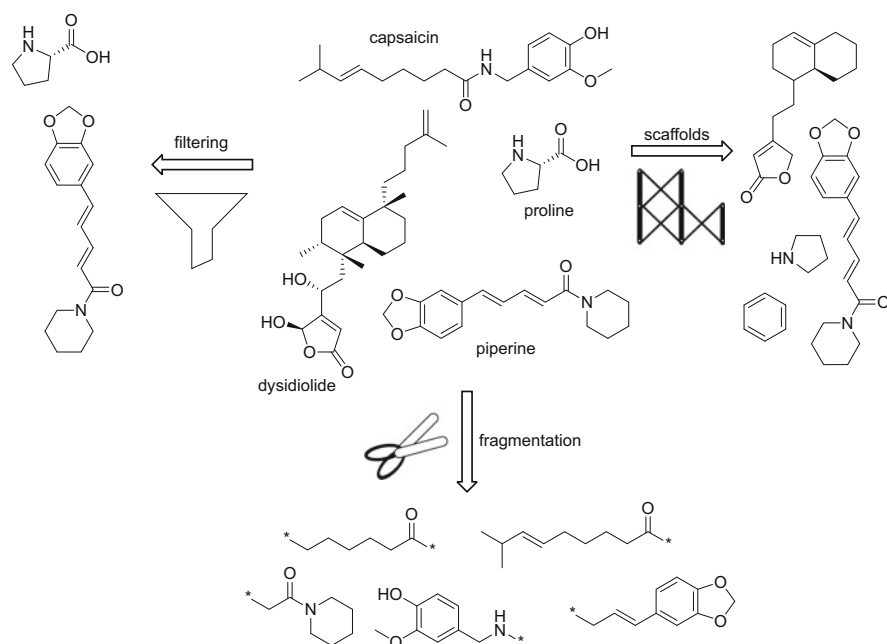


Fig. 1 Schematic on fragment generation. Shown is a natural product grouping consisting of proline, piperine, capsaicin, and dysidiolide. Fragment collections are shown that were generated relying on Murcko scaffolds [33], fragmentation through RECAP [15], and filtering (molecular weight <300). Asterisks indicate virtual attachment points generated through the in silico fragmentation. All procedures were implemented in RDKit [34]

determined by the desired goals of the specific project. This section outlines the most commonly used types of strategies and their various implementations.

2.1 Filtering

A straightforward approach to retrieve natural product fragments relies on filtering of natural product collections for their fragment-like sub-portion. Multiple studies have indicated that a large fraction of the currently known natural product space corresponds to low-molecular weight compounds [5, 17, 18, 35]. Thereby, molecular weight thresholds, commonly with a maximum allowed weight of around 300 Dalton, enable the rapid retrieval of the natural products that appear to be fragment-sized [8, 17]. More complex filtering strategies can ensure that the retrieved structures also fulfill other properties that are relevant for their concurrent application and thereby ensure the utility of the structures retrieved.

Indeed, many other definitions of fragment-likeness exist [36] and commonly other filters can be applied. A classic filter for fragment-like structures is the “Rule of Three” [37], a younger cousin of the Lipinski Rule of Five [32], which restricts fragments to a maximal weight of 300 Dalton, $\text{clog}P < 3$, up to three hydrogen bond donors (HBD), and up to three hydrogen bond acceptors (HBA). Generally speaking, the inclusion of pharmacophoric features (HBA and HBD, aromatic rings), measures of solubility ($\text{clog}P$, polar surface area), or measures of molecular complexity (number of rotatable bonds, number of rings, number of heteroatoms) can be rapidly derived from the molecular structures [38–40] and ensure that the extracted fragment collection fulfills the necessary criteria. Commonly, upper bounds on these properties ensure that the extracted structures are sufficiently simple and fragment-like [11]. It has also proven useful to establish lower bounds on properties such as the molecular weight, a minimum number of heavy atoms, or a sufficient molecular complexity to ensure that no trivial metabolites with limited pharmacophoric relevance are included in the natural fragment collection [17, 18].

Estimates of the fragment-like fraction of natural product databases range anywhere between 10 and 30% depending on both the investigated database and the exact definition of acceptable properties of the extracted fragments [5, 17, 18, 41]. While this number might be considered low, the vast size of many available natural product datasets [5, 26, 42, 43] usually still warrants a sufficient number of diverse fragment-sized structures for further applications [5, 8, 17]. Importantly, in absence of any computational structure generation or modification, this strategy promises to exclusively provide fragments that can be found in Nature and thereby might be available through isolation and, even more importantly, are more likely to constitute stable and biologically meaningful chemical matter [4].

2.2 *Virtual Fragmentation*

Instead of relying on smaller, fragment-like natural products already contained in natural product resources, *in silico* procedures to generate fragments can be harnessed to split larger natural products into smaller, fragment-like entities [7, 8, 15, 35]. Different strategies exist to virtually fragmentize larger structures into fragments, relying on distinct algorithms to ensure splitting of specific bonds or generating fragments of a specific size and character. Thereby, multiple fragments per natural product are generated, resulting in a potentially large set of virtual fragments that inherently contain natural product-like substructures [44]. A straightforward implementation of such fragmentation schemes can involve splitting of natural products at certain types of bonds, for example, all acyclic rotatable bonds [45]. Similarly, approaches to split up compounds into substructures of a specific size have been developed and applied to natural products [46, 47]. Following such strategies, exhaustive sets of substructures can be generated rapidly simply through analysis of the molecular graph and its partitions. However, many such generated fragments might be of limited chemical or biological relevance given their potential artificial nature and the lack of chemical reasoning for such partitions.

More advanced fragmentation schemes have been developed to ensure synthesis tractability, chemical stability, or other chemical and biological properties of the fragments. To this end, strategies have been implemented to virtually break bonds while considering their chemical context. Most prominently, the RECAP algorithm has implemented retrosynthetic fragmentation rules according to domain knowledge regarding bonds that can be easily and efficiently (re-)generated using established organic chemistry protocols [15]. In spite of not being specifically designed for natural product structures, the RECAP algorithm has been employed successfully to multiple natural product databases for the generation of large sets of natural fragments [7, 35, 41, 48]. Many publicly or commercially available cheminformatic software collections such as RDKit [34] and the Molecular Operating Environment [49] have re-implemented the RECAP algorithm with sufficient performance to process large natural product collections [7].

Many of the state-of-the-art *de novo* design implementations utilize virtual reactions. These construct novel compounds from commercially available building blocks according to hard-coded reaction schemes [50]. Pivoting this idea for fragmentation, reaction-based cleavage strategies have been implemented for natural products, for example, virtually performing reactions such as hydrolysis and ozonolysis [8]. These approaches might further the context-dependent analysis of cleavable chemical bonds and thereby generate virtual natural fragment-like collections with even further improved chemical relevance and a synthesis blueprint for their generation. Similarly, implementations of chemical reactions to introduce modifications in natural fragments can further diversify or stabilize the fragments of a virtual collection [4, 8]. While most of these approaches currently rely on hard-coded reaction schemes, novel approaches are emerging that can automatically extract such rules and further the applicability of computational fragmentation [51, 52].

Irrespective of the details of the implemented strategy, such in silico fragmentation approaches can rapidly and reproducibly provide large sets of natural fragments with novel chemical structures [7, 8, 48]. If virtual chemical reactions are utilized for fragment generation, these reaction schemes might provide blueprints for researchers to generate fragment-like molecules from a larger natural product serving as the template in the fragmentation. Even more importantly, virtual fragmentation is often implemented in such a way that it generates smaller structures with reactive handles that can serve subsequently as attachment points for further derivation and optimization [4] (cf. Fig. 1).

2.3 Scaffolds

Scaffold extraction can be regarded as a special case of fragmentation, whereas the molecular graph is not partitioned into smaller substructures but instead stripped into its central core structure by removing acyclic substituents [9]. While generally only one scaffold can be extracted per compound, the exact scaffold definitions and thereby the extracted molecular framework vary. The differences in the definitions concern whether the type of atoms or bonds are considered and whether adjacently connected heteroatoms are included in the scaffold definition [53, 54].

At the most abstract end of this spectrum are reduced molecular frameworks that ignore atom and bond types and even the size and types of rings within a scaffold [55, 56]. Molecular frameworks offer a mid-level of abstraction, representing full molecular graph-like structures utilizing all chemical bonds without atom and bond-type information. On the other end of this spectrum are decorated molecular substructures with fully defined atoms and bonds and the inclusion of specific adjacent heteroatoms [57].

Natural products, with their more common occurrence of fused-ring systems, will generate different reduced frameworks compared to synthetic compounds [57–59]. Full appreciation of the complexity and novelty of natural product scaffolds can only be achieved when including information on the distinct heteroatoms included or preservation of stereochemistry [60]. It is therefore not surprising that a large number of successful projects extract natural product scaffolds that consider atom-types, shortened side-chains, bond order, or chirality in their scaffold definitions [4, 9, 16].

A pioneering augmentation of scaffold generation and analysis is their hierarchical clustering based on substructure relationships to provide a structural classification of natural products (SCONP) [9]. Such hierarchical graphs of natural fragments enable tracking of chemical substructures and their impact on biological activity. This can aid significantly in the identification of lead fragments for further biological optimization [16]. Therefore, this concept has been implemented in the open-source software tool Scaffold Hunter to enable researchers to perform their own scaffold analysis [19]. Such approaches attest to scaffolds as some of the most useful natural fragment resource for further downstream analysis or as starting points for focused collection development (Table 1) [2, 9, 16].

Table 1 Number of natural fragments that can be generated from natural product databases relying on different computational strategies

Database	Fragment-sized	Scaffolds	RECAP
DNP	64,650	2,10,213	1,37,12,533
TCM@Taiwan	10,023	58,802	6,68,402
NuBBE	712	2218	21,892
AfroDB	191	954	3048
DMNP	≈17,941	28,833	17,77,882
DTNP	≈75,804	1,21,975	69,15,803

Data for TCM@Taiwan [43], NuBBE [61], and AfroDB [62] was analyzed in RDKit [34]. Data for DNP [42] were extracted from Reker et al. [7] and Rodrigues et al. [5]. Data for DMNP [63] and DTNP [64] were extracted from Shang et al. [35], whereas the number of fragment-sized compounds was estimated at half of the collection size given their reported median molecular weight

3 Properties of Natural Product Fragments

Although the exact properties of natural fragments vary widely and are influenced by the approach taken to extract them (Fig. 2), there are some general trends observable for natural fragments that render them a particularly useful and unique resource of chemical matter.

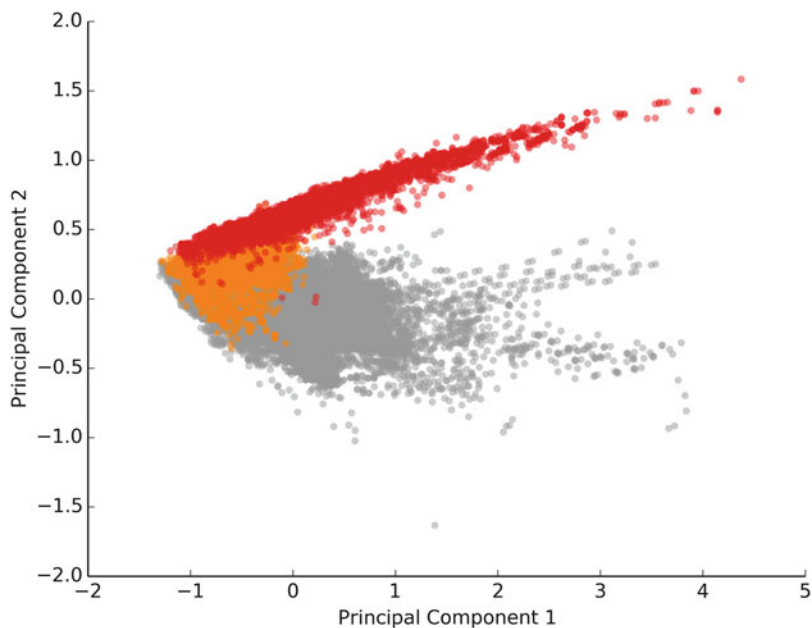


Fig. 2 Principal component analysis of standardized physicochemical properties of fragment-sized natural products (orange), scaffolds (red), and RECAP fragments (gray) from the TCM database. [43] Properties were calculated in RDKit [34], analyzed in KNIME [65], and visualized using Python and Inkscape (arbitrary units)

3.1 *Chemical and Physical Properties*

As expected, natural product fragments are more three-dimensional compared to flat synthetic fragments [4, 8] and rich in sp^3 -configured and chiral centers [5, 8, 66]. They also differ in their chemical composition and contain more oxygen, less nitrogen, and more aliphatic rings compared to synthetic screening collections [4, 12, 66], with an average number of three rings per fragment [9]. Some studies have investigated other properties relevant for biological activity and have shown a higher propensity for pharmacophoric features such as HBD and HBA and lower numbers of rotatable bonds [17], which has supported their perception as privileged structures with an increased potential to interact with a wider range of different biological targets compared to synthetic compound collections and their larger, complex natural counterparts [5, 7].

Chemography of natural products, their fragments, and drugs allows charting the general differences between compounds from these sources in terms of their physicochemical properties or chemical structures [67–70]. Generally speaking, researchers have found that natural product-derived fragments using various fragmentation sources generate good representative structures of the larger natural product collections in terms of fully spanning the natural product space [4, 57] while they show disparate placement compared to synthetic fragments [4] in terms of physicochemical properties. Pharmacophoric and structural assessments have placed natural fragments at the interface between synthetic bioactive compounds and complex natural products [7, 58].

Accordingly, less “Rule of Five” [32] violations are observed for natural fragments compared to natural products [18, 71], which might not be too surprising given that these rules include thresholds that depend heavily on the size of the investigated molecule and therefore are commonly conformed to by natural product fragments. Fascinatingly, natural products in general often violate the “Rule of Five” [12, 72], such that fragmentation might be regarded as a transformation of natural product space into “Rule of Five” compliant areas [73]. This is fully in line with the observation that the regions of physicochemical space that are populated exclusively by natural products but not drugs are mostly spanned by larger, more complex natural product structures [7, 57]. Thereby, fragments enable a more drug-like handle to natural product space [2]. This can be further utilized since their decoration with classic medicinal chemistry side chains enables populating drug-like spaces but with innovative scaffolds [57] with potentially superior properties such as higher three-dimensionality [8].

3.2 *Spatial Properties*

To assess the size and shape of natural product fragments, the three-dimensional conformation can be predicted from the two-dimensional molecular graph

[74]. Conformation prediction of natural products is often challenging given their large size, complex structures, and common occurrence of chiral centers and macrocycles [75, 76]. Therefore, natural product research often harnesses advanced methods relying on experimental methods such as nuclear magnetic resonance measurements or machine learning predictions [77, 78], thereby limiting the application to only select natural products with available experimental data or sufficient interest to warrant the necessary experimental or computational resources. Since fragment-like natural products are smaller and populate a more restricted conformational search space, their conformations might be estimated more rapidly and more accurately [36]. This suggests that natural product fragments and their derivatives might enable researchers to study natural products in three dimensions with reduced need for advanced conformation estimations to identify targets via pharmacophore searching or rationalizing binding modes through docking [9, 21, 22, 79, 80]. On a more general level, such three-dimensional structure prediction tools have also been employed to whole collections of natural product fragments to calculate the distribution of their volumes: this has charted most natural fragments within the range of 100 to 500 Å³. This renders them similar in size compared to volumes calculated for approved drugs and currently explored protein pocket cavities [9, 16, 35], further attesting to their utility for drug discovery. Similarly, the shape of natural product fragments has been investigated through their principal moments of inertia [81] and it was concluded that natural product fragments provide a wide range of different shapes and, most notably, are less “flat” compared to other compound databases, including the complex natural products stored in the Dictionary of Natural Products [42] as well as various synthetic fragment collections [8, 17].

3.3 *Natural Fragments from Different Sources*

Natural products and their fragments from different origins can vary drastically in their chemical structures and properties [12]. By selectively analyzing natural products from different sources, their properties can be compared to identify potentially helpful trends in physical or chemical differences between structures produced by distinct organisms or within specific environments or locations [35, 41, 66, 82]. A handful of studies have started to chart and compare natural products from different sources. For example, marine products seem to offer a larger variety of different substructures, while terrestrial products appear to borrow more frequently from similar substructures [35]. It has been suggested that natural products originating from fungi might represent distinct natural product properties without deviating too much from the drug-like space in terms of physicochemical descriptions [57]. Ertl and Schuffenhauer specifically investigated unusual chemical structures populating natural products from different sources and found that plant-based natural products contain more fused carbocycles compared to natural products from other sources [83]. Other studies have found that arenes, while dominating plant- and marine-derived natural products, seem to be almost completely absent from bacterial

organisms [66]. Bacterial metabolites may also be sulfur-containing natural products [83] and marine natural products contain more oxygen compared to terrestrial natural products [35]. Isolated studies have started to draw conclusions from observed chemical structures and the implications for their occurrence in certain natural products: for example, marine natural product repositories contain more hydrophobic compounds and a lower number of ester bonds [35]. This could point potentially at evolutionary forces selecting for organisms that are more adapted to their marine environment by producing compounds with lower risks of losing metabolites to the aqueous environment surrounding them as well as spontaneous hydrolysis [35, 83]. Such insights are transferable into the fragment space of natural products and can be helpful in compound collection design when certain physico-chemical properties are of importance. Relying on selected organisms or origins might enable compound pools to be steered in the desired direction. Furthermore, through such in-depth analysis, cheminformatic research potentially may assist fundamental research in metabolomics to better understand specific organisms or microenvironments through their small molecular armamentarium [84, 85].

3.4 Commercial Availability and Synthetic Tractability

Although fragment-sized natural products make up only around a third of the known natural product structures [5], they constitute the bulk of commercially available and hence easily-accessible natural compounds for drug discovery and biotechnology applications [4, 26]. Natural product fragments are often easier to synthesize compared to their complex, larger counterparts [4, 73]. Computational assessments of synthesizability as well as computational retrosynthesis planning potentially can aid at prioritizing natural fragments and synthesis pathways [15, 51, 86]. However, most of these tools were not designed to be specifically applicable to natural products and therefore might need to be augmented to enable their straightforward application to natural fragments. It has also been shown that, even in cases where the fragments are not readily available or synthesizable, they can often be represented through commercially available or easily synthesizable analogs [4] (Fig. 3). This can be achieved by employing classical ligand-based similarity and virtual screening approaches to search for similar fragments [87, 89]. Alternatively, clustering of fragment spaces enables partitioning the chemical fragment universe into regions of high similarity to substitute critical fragments with other co-clustered representatives [4, 7]. Extending this concept even further and relying on hierarchical clustering techniques, Koch et al. showed that constructing graphs through chemical substructure relationships can build data structures to enable such simplifications [16]. Applying this concept to scaffolds, such a graph can be built iteratively by populating the network with all possible scaffolds. Subsequently, two nodes are connected if the respective scaffolds can be transformed into each other by either removing or adding one ring structure. Thereby, a graph is generated based on a special case of a substructure relationship. Traversing this graph [19] can inform chemical derivatization and simplifications

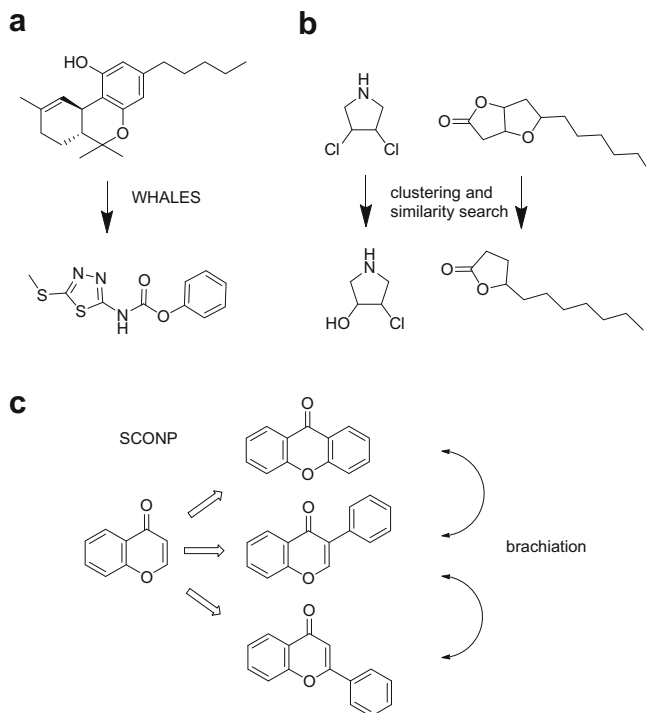


Fig. 3 Identifying accessible derivatives of natural products, e.g., through molecular similarity assessments, using, for example, the WHALES descriptor [87] (a), or through clustering and subsequent selection of representative structures via chemical structure similarity (b) [4]. Alternatively, hierarchical graphs of scaffolds can be utilized to identify related structures following the SCONP concept (c) [9, 88]

into smaller scaffolds with lower complexity and sufficient similarity to enable tackling of the same biological target [90]. If a substructure of the scaffold does not warrant a sufficient simplification, such graphs also enable “brachiating” into neighboring branches with potentially simplified chemical structures but retained biological activity [9]. In such advanced simplification approaches, it is important to keep in mind that too aggressive derivatization or simplification can lead to loss of the biological activity or other desired properties of the investigated natural product fragment [9, 91]. Prediction of the biological and physicochemical properties of the *in silico*-derivatized structures can enable researchers to monitor the expected behavior of the novel structures and guide the structural modifications [5, 22, 25, 92–95]. Conversely, the originally investigated fragments might be unstable or contain reactive substructures such as enamines or Michael acceptors, and slight modifications can easily eliminate such structures [4, 8, 20]. The prediction of reactive substructure or other liabilities can assist in identifying problematic fragments and speed up this process [96, 97].

3.5 *Problematic Natural Fragments*

An increasing body of literature has served to warn of false-positive assay results, i.e., compounds that elicit a positive readout in spite of not showing the actually desired biological activity [97, 98]. This behavior can originate from various underlying causes, many of which relate to physicochemical properties of the investigated compound such as reactivity, quenching, membrane interactions, fluorescence, or colloidal aggregation [97–99]. Natural products and their fragments can potentially contain substructures that can elicit such effects *in vitro* even if such effects might be masked in their natural, biological (micro)environments [12]. This hints at the necessity to identify problematic structures in natural product collections and for fragment-like hit compounds to avoid hunting artifactual biological *in vitro* results [99].

Cheminformatics efforts to design automated pattern-recognition systems to filter such potentially problematic compounds have led to the development of multiple substructure-based filtering lists that flag chemistry with motifs associated with false-positive results for subsequent validation or elimination [97, 100, 101]. Although such methods have been designed based on screening data or with synthetic molecular probes in mind, it has been shown that such false-positive behavior can occur among natural products in general and their fragment-like portion in particular [101, 102]. False-positive results from natural products might be common, as indicated by multiple case studies and in-depth analysis of commonly used natural fragments [101, 102]. For example, the discovery of viable inhibitors of indoleamine 2,3-dioxygenase 1 (IDO1) has been fueled by fragment-like natural products such as β -carboline or galanal, but it has been suggested that potential false-positive readouts might be at play for such compounds [103]. Similar arguments have been made for fragments such as thymoquinone in the context of anti-protozoal agents [104] and plumbagin as a histone acetyltransferase inhibitor [105]. On a larger scale, statistical data analysis of natural product assay data has been conducted and suggests that a large fraction of the acquired positive assay results might stem from causes other than useful biological activity [99].

To counter such effects, researchers have designed taboo lists of chemical substructures that are linked to various artifactual readouts [97]. Many natural products contain such critical chemical substructures [12] recognized by filtering rules such as PAINS [2, 99, 101] or ALARM NMR [100, 106, 107]. In fact, it appears that natural products are particularly prone to contain such critical substructures recognized by automated filtering rules compared to synthetic compounds and approved drugs [5]. For example, around 40% of natural products currently studied in the context of their antiprotozoal activity contain PAINS substructures [104] and up to 65% of natural products from the commercial MicroSource collection are flagged according to the ALARM NMR filters [107]. While, to the best of our knowledge, no large-scale analysis on flagging of natural fragments has been conducted, it is fair to assume that fragmentation or selection of potentially problematic natural structures would transfer such liabilities into fragment-based collections and pipelines

Table 2 Percentages of natural product fragments flagged by PAINS [96] and ALARM NMR [100] filtering rules for select datasets from Table 1 [43, 61, 62]

Database	Fragmentation	ALARM NMR/%	PAINS/%
TCM@Taiwan	Fragment-sized	34.2	0.7
	Scaffolds	38.8	0.9
	RECAP	24.6	0.6
NuBBE	Fragment-sized	28.2	0.8
	Scaffolds	25.4	0.5
	RECAP	24.0	0.5
AfroDB	Fragment-sized	69.6	16.8
	Scaffolds	41.2	3.1
	RECAP	50.9	8.1

Flagging was performed using the webserver made available through the Division of Biocomputing, Department of Biochemistry and Molecular Biology, University of New Mexico, Albuquerque, NM (<http://pasilla.health.unm.edu/tomcat/biocomp/smartsfilter>)

[108]. Furthermore, specific investigations of selected fragment-sized natural products have highlighted cases of such naturally occurring structures to be flagged by multiple different false-positive detection methods (Table 2) [103].

Colloidal aggregation has been suggested as the single largest reason for a compound to elicit artifactual, false-positive assay readouts in screening assays [97, 109]. Indeed, many fragment-sized natural products such as physcion and equol have been shown to form colloidal aggregates that can sequester proteins and thereby interfere with biochemical assay readouts [99, 102, 110]. A high $\log P$ for many natural products [58, 60] is an indicator that natural products and their fragments might possess such aggregation propensity [109]. More accurate and fine-grained computational prediction models exist that anticipate whether a compound aggregates from its molecular structure and physicochemical properties [109, 111, 112]. Such models might be applied fruitfully to get more accurate estimates on which natural product fragments form colloidal aggregates. However, given that such models rely on molecular data mostly derived from synthetic screening compounds [112], the discrepancy between natural compounds and the training data in terms of molecular properties and structures hint at natural products potentially lying outside of the applicability domain of such models [7, 113, 114]. An in-depth evaluation will be necessary to understand whether the colloidal aggregation of natural products can be studied by relying on data derived from synthetic screening compounds or whether specifically tailored machine learning tools will be necessary to accurately delineate the potentially common aggregation behavior of natural products [102].

Furthermore, fragment-like natural compounds such as genistein and capsaicin potentially can interact with and modulate lipid bilayer properties and thereby cause false-positive readouts in cell-based screening assays [115]. To rapidly decode such effects and flag compounds that potentially exert such behavior, computationally accessible properties such as lipophilicity [116], charge [117], and amphiphilicity [115, 118] can enable model development to predict the ability of natural fragments

to interact with or modulate lipid bilayer properties. Molecular dynamic simulations are another tool that can potentially anticipate such effects and identify the natural fragments that can cause this type of behavior [115, 117, 119].

Not all natural fragments that are flagged by such methods need to be blindly eliminated [120]. Many liabilities will only be relevant in specific screening contexts [121, 122], which has led to major criticism against the blind application of the aforementioned prediction models and flagging lists to eliminate compounds from screening collections [98, 120]. Indeed, many safe and clinically effective medications have been shown to be flagged by various computational false-positive detection methods [123, 124]. Therefore, if sufficient caution is taken and validations and counter-screens are established, even apparently problematic structures might fuel successful drug discovery and development pipelines. In the future, augmented prediction methodologies could enable more fine-grained analysis of contextual assay results of natural product fragments [106]. Other studies have shown how natural fragments containing problematic structures constituted initial hits and were subsequently derivatized during optimization to eliminate the liability for downstream validations [8] or how potentially pernicious substructures might not be liabilities in their specific natural product context [5, 17]. Utilizing computational prediction models to anticipate context-specific liabilities, as well as the establishment of automated molecular design for the derivation of natural product fragments will automate such processes in the future.

4 Applications of Natural Product Fragments

Given the aforementioned advantageous properties of natural fragments, they have been suggested as representing innovative starting points for drug discovery and chemical biology [4, 5, 13]. Importantly, such efforts are most productive if the polypharmacological properties of the natural product under investigation is known and derivatives can easily be made to fine-tune biological and physicochemical properties of the molecular probes. As described in this section, computational tools can assist in predicting these properties, generating focused collections of derivatives and mimetics of natural compounds, as well as utilizing them for statistical analysis of natural product-likeness to assess the utility of a compound or molecular collection.

4.1 *Predicting Biomacromolecular Targets of Natural Fragments*

Natural fragments with known polypharmacological profiles may be regarded as most useful starting points for drug discovery campaigns and chemical probe

development since candidate structures will already possess the desired activity and potential known off-targets can be avoided [5]. Unfortunately, the biological effects of a vast majority of natural products and their fragments is currently not known [6, 7, 9]. In silico target-interference methods represent easily deployable prediction tools to anticipate the biomacromolecular receptors targeted by natural products [5]. However, classical target prediction methods usually underperform for natural product fragments given the stark difference in chemical structure of natural products and their fragments to the synthetic compounds that populate (training) databases of ligand-target interactions [4, 5, 18]. Therefore, with few exceptions [125, 126], most of the advanced and well-validated target prediction technologies, which are largely based on applying the chemical similarity principle [127] to chemical substructure descriptions [128, 129], underperform at identifying targets for natural product fragments compared to their impressive success reported for the target identification of drugs and synthetic compounds [95, 130–133]. Therefore, in the context of natural product fragments, researchers have employed or designed target prediction methods that generalize from the underlying chemical substructure and instead directly or indirectly quantify the pharmacophoric potential of natural product fragments [5].

For example, Rollinger et al. have used 2208 three-dimensional pharmacophore models to screen a collection of 16 fragment-like secondary metabolites isolated from *Ruta graveolens* and found between ten and 287 confident predictions per natural fragment [79]. In prospective experiments, arborinine was validated successfully as an acetylcholinesterase inhibitor ($IC_{50} = 34.7 \pm 7.1 \mu M$). Their model for binding the cannabinoid-2 receptor was based on five selective agonists. The only confident prediction of this model was rutamarin. Indeed, rutamarin was the only metabolite that showed ligand displacement with a K_i of $7.4 \pm 0.6 \mu M$ [79]. These data suggest that such models are not only able to correctly identify inhibitors but also are robust in recognizing true negatives—although further testing will need to statistically validate these results [79]. Follow-up research has led to the identification of acetylcholinesterase inhibitors among morphinans and isoquinolines [134] as well as partial agonists of proliferator-activated receptor gamma among neolignans [80, 135].

Instead of deriving a pharmacophore model for a protein target of interest, researchers have also successfully employed docking strategies to assess the potential of a natural fragment to bind a pocket of a target protein [136–138]. For example, Lanz and Riedl probed the S1'-binding site of matrix metalloproteinase 13 [22]. They identified uracil as a natural fragment with a unique binding mode that was further optimized into nanomolar inhibitors with impressive selectivity over other matrix metalloproteinase subtypes. In a broader screen against 400 proteins, Bernard and colleagues identified peroxisome proliferator-activated receptor gamma and cyclooxygenase-2 as targets of the coumarin derivative meranzin [139]. Both pharmacophore models and computational docking strategies have shown impressive results for identifying biomacromolecular targets of natural product fragments [5], but require good estimates of the conformational space of natural fragments and their stereochemistry, which are not always available.

Other prediction technologies enable target identification for natural products while circumventing the challenges originating from the differences in chemical structure compared to synthetic screening collections or unknown conformations [5]. For example, researchers have employed productively biological fingerprints as an alternative strategy to compare molecules and predict their targets [140]. In brief, the underlying assumption is that if two compounds have shown similar activities for some biological targets, it is likely that they will behave similarly when tested against other targets. This can be employed for target prediction if one of the compounds has been tested against targets against which the other compound is yet to be tested. Wassermann et al. have shown how this approach predicts more targets for natural products compared to chemical fingerprint-based approaches and have used this concept to identify vascular endothelial growth factor receptor 2 as a target for fisetin with an IC_{50} of 230 nM, which might aid explaining the antiangiogenic effects of this fragment-like flavonoid [141]. However, this successful strategy is exclusively applicable to natural products that have been screened for biological activity previously. A majority of natural product fragments have not been investigated yet or the results have not been made publicly available [6, 7], highlighting the need for additional technologies that can predict targets of natural fragments exclusively from their structures without the need for conformational sampling or previous biological screening.

In an effort to design a prediction technology specifically focusing on its ability to predict targets for novel chemical structures, Reker et al. have designed the SPiDER method [25]. The method circumvents the problem of predicting targets for chemicals with unusual or previously underexplored chemical substructures by explicitly employing “fuzzy” descriptors that enable relating chemicals through their two-dimensional graph structure via pharmacophore correlations (CATS2 descriptor) [142, 143] and physicochemical properties [39, 49]. The method relies on self-organizing maps as a clustering approach [144] to define local regions (Voronoi fields) of equivalent biological activity [145, 146]. Confidence scores for every prediction are derived from statistical interpretation of molecular similarities to enable prioritization of the most meaningful target hypotheses [25]. It was realized that this workflow not only is successful at identifying targets for novel, de novo-designed synthetic compounds [25, 147] but specifically excels at predicting targets of natural products [5, 7, 21, 148]. Natural fragments, in particular, appear to show the highest number of confident SPiDER predictions [5, 7], highlighting the ability of the SPiDER algorithm to predict new targets for this important compound class. Indeed, SPiDER has been utilized to identify biomacromolecular targets for various naturally occurring fragments such as β -lapachone [149], graveolinine, isomacroidin, DL-goitrin [21], sparteine [5], valerianic acid, isopimaric acid, and dehydroabietic acid [20] (Fig. 4). While these and other studies [7, 148] highlight the power of ligand-based target prediction methods such as SPiDER to identify the targets of natural fragments, these studies also provide powerful insights into how these methods can be used in concert with molecular docking [21], molecular similarity assessments [20], and orthogonal machine learning technology [149] to fuse multiple prediction methodologies for further improved predictive confidence or to enable

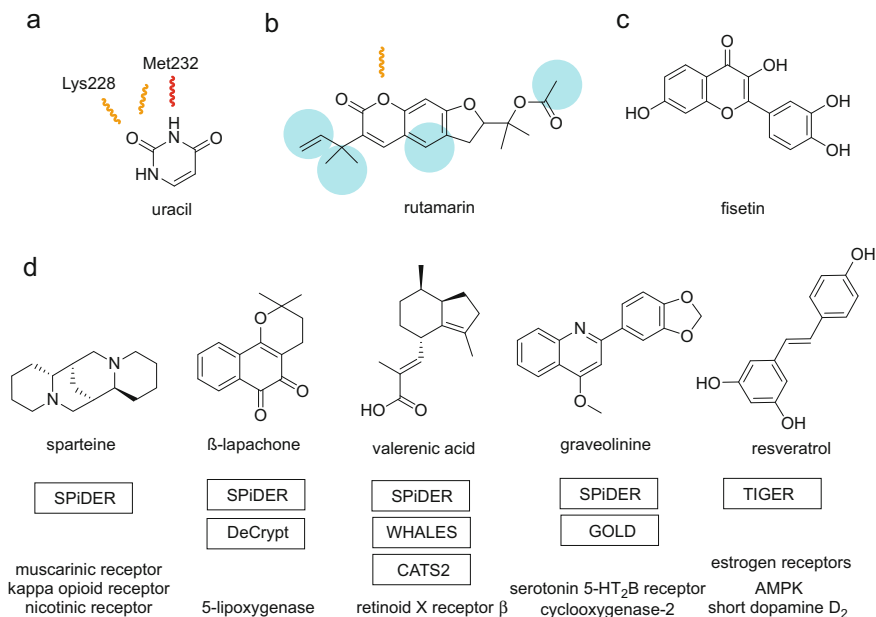


Fig. 4 Predicting biological targets of natural fragments. **(a)** Prediction of the binding mode of uracil in the S1'-binding site of matrix metalloproteinase 13 [22]. Orange, hydrogen bond acceptor; red, hydrogen bond donor. **(b)** Pharmacophore-based identification of rutamarin as cannabinoid-2 binder [76]. Orange, hydrogen bond acceptor, cyan, hydrophobic interaction. **(c)** Biological activity fingerprints enabled the identification of endothelial growth factor receptor 2 as a target of fisetin [133]. **(d)** Selection of different fragments for which the targets were identified through ligand-based target prediction technology [5, 20, 21, 149, 150]. The target prediction method utilized for the individual study is highlighted in the boxes underneath the structure and the identified targets are annotated at the bottom

additional hit rationalization. Similarly, conceptually related target prediction methods such as TIGER show promise to further the computational toolset for polypharmacological prediction of natural product fragments such as resveratrol [150] that can be easily derivatized for further structure-activity relationship studies and in vivo applications [151].

Although multiple striking examples exist of utilizing advanced target prediction technology to predict targets of large and structurally intricate natural products [7, 125, 148, 152, 153], it seems that fragment-like natural products are more computationally relatable to the available screening data [2, 7, 57] and therefore lead to more confident predictions [5, 7, 21, 107]. Some investigations suggest that this trend might also be true for other target prediction methodologies such as computational docking, where fragment-like entities lead to higher scores or improved retrieval of correct binding modes [36], which might be connected to the problem of conformational sampling of complex natural product structures [78].

Taken together, it appears that fragment-like natural products are exquisitely positioned to provide starting points for drug and chemical tool development to

modulate the activity of their anticipated targets [13]. For example, sparteine is a natural fragment that has been studied extensively and computational methods such as ligand-based target prediction and clustering-based diversity selection have identified biomacromolecular targets spanning different protein families such as p38 α MAP kinase [4], muscarinic and nicotinic receptors [5, 154], and the kappa opioid receptor [5]. Such initial hits can then be further optimized by derivatizing the fragment and adding additional chemical functionality: sparteine derivatives were further functionalized with a primary amine that enabled improved p38 α MAP kinase activity from additional polar interactions as indicated by molecular modeling [4].

4.2 Collection Design

Natural product fragments are believed to coalesce the advantages of fragment-based drug discovery and biologically privileged natural product structures [13]. Therefore, general-purpose screening collections of natural fragments with diverse structures have been harnessed to generate compound collections that could provide novel hits at improved rates with better selectivity compared to classical synthetic fragment sets [8, 11, 17, 155]. For example, Over et al. [4] used scaffold-based fragmentation on the Dictionary of Natural Products [42] and subsequently filtered for fragment-like structures without reactive groups. Clustering-based diversity selection [156] and identification of commercially available cluster representatives (cf. Fig. 3) lead to the assembly of a screening collection that was employed successfully to identify novel, allosteric ligands of p38 α MAP kinase as well as phosphatase inhibitors [4]. Similarly, Quinn and colleagues [17] used filtering directly on the Dictionary of Natural Products [42] to arrive at a small, diverse collection of naturally occurring, three-dimensional fragments that were successfully isolated or purchased. This compound collection was then screened in parallel for their potential to bind to multiple malarial protein targets as well as phenotypically for their activity against asexual intraerythrocytic blood stage *Plasmodium falciparum* 3D7 parasites. Analyzing these data in parallel enabled the target identification of 31 relevant antimalarial targets as well as the generation of 79 innovative hit structures for further optimization [17]. Such efforts attest to the enormous potential of diverse, target-agnostic screening collections composed of natural product fragments to fuel fragment-based discovery efforts against targets from various protein families using various assay technology as well as validating their utility in phenotypic screens.

Instead of generating natural fragment sets from whole natural product collections [4, 17], molecular series derived from one specific natural product core can provide focused collections with privileged and novel structures [12, 16]. Especially when the biomacromolecular targets of the template natural product are known or predicted [9, 149, 157], the derived structures often inherit the template's polypharmacological profile—leading to dramatically increased hit rates for the

focused natural fragment collections on the target of interest [9, 158]. Furthermore, if the template fragments were generated *in silico* through virtual fragmentation and (retro)synthesis approaches, their attachment points might constitute useful chemical handles to add side chains with additional pharmacophore functionality while preserving the original scaffold and shape [4] (cf. Fig. 1). Thereby, potentially inaccessible natural products can serve as templates for synthetically tractable sets of derivatives [91]. Waldmann and colleagues have pioneered and validated this concept as biologically oriented synthesis (BIOS) [16]. Relying on and derivatizing core structures originating from fragments that were computationally extracted from natural ligands of an enzyme of interest or other known or predicted inhibitors, focused screening collections can be generated with high hit rates and, even more importantly, an often improved selectivity to structurally related protein targets compared to other screening approaches [9, 90, 158].

Fascinatingly, such efforts can be performed phenotypically without necessarily understanding the exact mechanism of action of the template compounds. For example, in an effort to identify molecular tools with neurotrophic activity, Schröder et al. relied on BIOS to simplify *N*-deoxymilitarinone A (Fig. 5), a fungal metabolite causing neurite outgrowth in PC-12 cells [90]. Previous research had shown that this effect is relatively robust toward simplifying the natural product and its side chains have minor impact on its neurotrophic activity [159, 160], suggesting a pivotal role of the scaffold for activity. This motivated the generation of 59 compounds around the 4-hydroxy-2-pyridone scaffold and its 2,4-dimethoxypyridine derivative [90]. The most active compound **11e** showed 74% neurite growth at 10 μ M compared to control. Interestingly, activity in the phenotypic screen of the focused collection could be correlated with MAP 4K4 activity, which potentially links this kinase to neurotrophic effects and proposes it as a target to combat neurodegenerative diseases [90].

An earlier study by Koch et al. set out to identify novel 11 β -hydroxysteroid dehydrogenase 1 inhibitors from the natural ligand glycyrrhetic acid [9]. The complex pentacyclic scaffold was simplified into a two-ring system following the SCONP hierarchical clustering of scaffold structures [9]. In an additional step of collection development, the scaffold was subsequently substituted by a more stable derivative with endocyclic double bond through “brachiation” within the scaffold tree [161]. Such a horizontal shift from one arm of the scaffold tree into another can enable transformations of target fragments into more suitable structures in terms of advantageous physicochemical properties or improved chemical tractability for derivation and collection design. However, this comes at the risk of losing the associated pharmacological effect of the original natural product through deviating too much from the privileged scaffold arrangement [9, 91]. In this specific case, further confidence in the applied modification was drawn from the fact that the aspired fragment corresponds to the scaffold of the natural product dysidiolide, which is an inhibitor of Cdc25A phosphatase [162]. Since Cdc25A phosphatase is structurally related to the 11 β -hydroxysteroid dehydrogenase 1 target protein according to the protein structure similarity clustering (PSSC) approach [158], it is likely that they share common inhibitors and that therefore scaffolds targeting

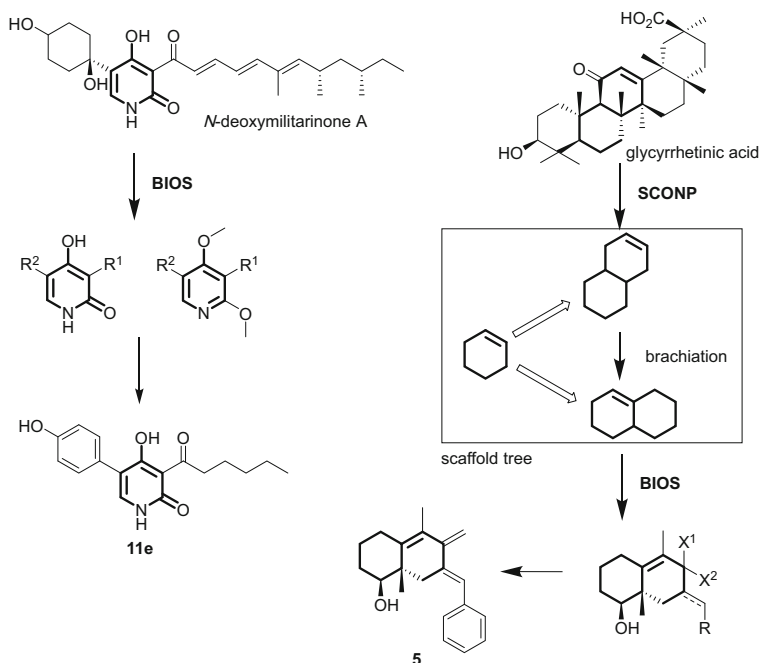


Fig. 5 Collection design using biology-oriented synthesis (BIOS). Natural ligands with known biological activities, such as *N*-deoxymilitarinone or glycyrrhetic acid, can be simplified into smaller scaffolds through traversing the structural classification of natural products (SCONP) graph. This leads to the identification of a core scaffold that can be used to generate focused collections of derivatives with a desired biological activity [9, 90]

Cdc25A phosphatase might also warrant 11 β -hydroxysteroid dehydrogenase 1 inhibitors. Indeed, a collection of 162 compounds derived from the octahydronaphthalene scaffold afforded 30 inhibitors of 11 β -hydroxysteroid dehydrogenase 1 and some of the hits revealed remarkable selectivity against 11 β -hydroxysteroid dehydrogenase 1 over 2 [9], further hinting at the potential of natural fragments to serve as starting points for highly selective lead structures and probes [5].

In these and other examples [16], BIOS has specifically excelled at addressing challenging targets where out-of-the-box fragment or screening collections might give unsatisfactory results [4]. This has been associated with the ability of such approaches to generate focused sets of compounds that inherit relevant physico-chemical properties and pharmacophores from the template natural product [4, 9, 16]. However, notwithstanding the impressive success rate of the BIOS approach, this hypothesis is not always correct, and depending on the chemistry employed, compounds with vastly different properties can emerge [9, 91]. In a fascinating meta-analysis, Pascolutti and Quinn [163] investigated the distribution of molecular weight and log*P* as well as the number of HBD, HBA, rotatable bonds, and rings in collections generated through derivatizing natural product templates. They found

that such generated collections can show markedly different ranges of properties and, most importantly, could be drastically different from the original natural product that was used as the template structure [163]. A close monitoring of the properties of collections generated, including their potential to trigger false-positive assay readouts (cf. Sect. 3.5) potentially can improve the quality of such generated collections and even further enhance success rates [5].

The impressive success of the BIOS approach is possible due to advanced chemical knowledge and manual labor to delineate stable core structures and suitable chemical routes for their derivation [16]. In an orthogonal and automated approach, computational de novo design can be installed to generate large collections of natural product mimetics autonomously [20, 58, 92, 164]. To this end, natural product fragments with known biological activities can fuel ligand-based de novo design algorithms [50] to derivatize them and build small collections of novel chemical structures with similar chemical features and biological activities [92].

Schneider and colleagues have pioneered both established as well as novel de novo design methods for the generation of collections of synthetically accessible natural product mimetics from natural product fragments [20, 92]. For example, the DOGS software [24] was harnessed to create a collection of synthetically accessible natural product mimetics with potential inhibitory effect against the retinoid X receptor [20]. DOGS implements a virtual synthesis algorithm that connects 25,144 commercially available chemical building blocks according to 58 chemical reaction principles to generate novel chemical matter with a suggested protocol for its synthesis [24]. Since the connection of all building blocks would lead to a combinatorial explosion of possibilities, the DOGS designs are iteratively guided by a graph kernel similarity towards a template structure [165]. The DOGS designs have been validated extensively to afford novel chemical matter with the desired activity against (patho)biologically relevant protein targets [166–172] and were most recently validated in the context of fragment-based drug discovery [173] and were used to identify mimetics of large, complex natural products [164]. In a consequent next step, Schneider and colleagues applied the concept to natural fragments as retinoid X receptor modulators, utilizing small natural compounds with known retinoid X receptor activity such as honokiol, drupanin, valerenic acid, isopimaric acid, and dehydroabietic acid as template structures (Fig. 6) [20]. These designs were prioritized further using a consensus of a CATS2 descriptor-based similarity assessment [142] as well as SPiDER target predictions [25]. This workflow led to the synthesis of six de novo-generated natural product mimetics, of which five showed the desired retinoid X receptor activity [20].

In a second study against the same target, the team added bigelovin to the natural product templates for a novel, generative deep-learning campaign [92]. Generative deep neural networks are the newest addition to the molecular design toolbox, teaching a machine molecular structure constraints by providing tens of thousands of valid chemical structures in text representation such as SMILES formats [174–176]. These neural networks can then sample novel text representations of molecules that translate into compounds with desired properties. To this end, the neural network is first trained to produce chemically meaningful text representations

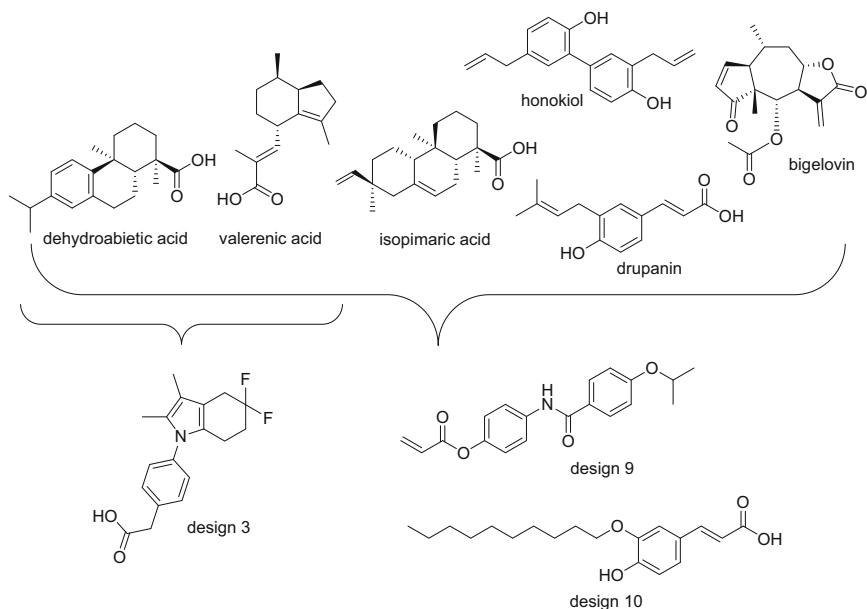


Fig. 6 De novo design of natural product-derived fragments for the generation of synthetically accessible mimetics. The natural fragments and inhibitors of retinoid X receptor, honokiol, drupanin, valeric acid, isopimaric acid, bigelovin, and dehydroabietic acid can be coupled to ligand-based de novo design software such as DOGS [24] or deep learning-based generative models [174] to identify novel chemical entities with natural product-likeness [44] and a desired biological activity [20, 92]

through a large corpus of chemical structures. Providing the six natural fragments as data for an additional round of training enables the fine-tuning of the model and bias the generation of chemical structures to natural mimetics with potential for the desired biological activity (“transfer learning”). Indeed, the fine-tuned model was able to create hundreds of chemically valid and novel structures that exhibited the desired natural product-likeness [44]. Around half of these designs were predicted as ligands of retinoid X receptor by orthogonal target prediction methodology relying on SPiDER predictions [25]. Further filtering of these positively predicted designs using WHALES molecular similarity assessment [87] and visual inspection for synthesizability lead to four chosen designs, of which two possessed retinoid X receptor activity with varying subtype selectivity [92].

Such de novo design campaigns are complementary to the previously mentioned BIOS approaches. While BIOS harnesses chemical expert knowledge to generate structurally related compound series employing the same or a similar privileged scaffold, automated molecular design can create large sets of novel structures that are chemically more different from the template but inherit natural product-likeness [44] and crucial pharmacological features from the template natural fragments [73, 92]. Given their impressive success in previous studies, the herein discussed

and orthogonal tools are likely to become essential parts of the drug and chemical probe discovery toolbox to provide novel and privileged compound collections from natural fragment templates [73].

4.3 Analysis of Natural Product-Likeness

Instead of using fragments directly as tools in computer-assisted drug discovery, researchers have utilized them as chemical patterns to quantify the similarity between two compounds or compound collections. Scaffolds in particular have been utilized as the indicators of structural novelty for collection design and compound development [18, 58, 132, 177]. Therefore, the concept of “scaffold hopping” has been coined by Schneider and colleagues to ascribe the capabilities of a method or the success of a project to identify new chemical entities [143, 167]. Conversely, in the context of natural product-inspired drug discovery research where scaffold similarity to a naturally occurring compound is desired, the detection of natural product fragments in a new lead structure or screening collection is a measure of desirability [44]. On a larger scale, fragmentation and structure matching can provide statistically quantifiable measures of “natural product-likeness” [5, 44]. For example, Hou and colleagues [35] analyzed the natural product-likeness of the Comprehensive Medicinal Chemistry Database and found that about 20% of the scaffolds in this dataset were present in the Terrestrial Natural Product Database [64] while only 10% could be found in the Dictionary of Marine Natural Products [63]. While this bias might be in parts explained through the smaller size of the Dictionary of Marine Natural Products (cf. Table 1), other effects such as the underrepresentation of marine natural products in historic drug discovery might help explain this effect as well [57].

To measure natural product-likeness of individual chemical structures, Rodrigues et al. have devised a scoring procedure that captures the count of natural product fragments occurring in one specific molecule normalized by its molecular weight [5]. Through this normalization, each score captures the relative frequency of natural product fragments found within one specific structure compared to the size of the molecule. Through applying this score to FDA-approved drugs, therapeutics that are more or less similar to natural products can be identified. Interestingly, while there were strong variations per year, a sustained occurrence of natural fragments in approved drugs was observed [5]. This is fully in line with previous observations of the relevance of natural products for drug discovery [3, 178].

Ertl et al. have devised the most commonly utilized score that specifically captures the occurrence of fragments from natural products that cannot be found in synthetic molecules [44]. Thereby, the score removes “background noise” fragments that can be found in either compound class. This score enabled the classification of natural products vs. synthetic molecules with higher enrichment compared to machine learning models based on physicochemical properties, thereby highlighting its utility to measure the natural product-likeness of a compound in terms of

chemical fragments. In a striking experiment, they compared the scores for natural products, synthetic compounds, and approved drugs: while natural products and synthetic compounds show disparate distributions, approved drugs show higher natural product-likeness, further attesting to the utility of using natural product fragments in the design of new therapeutics [179].

Accordingly, such measures of natural product-likeness act not only as a seismograph to measure the relevance of natural products for drug discovery over time [5], but have also been used to guide screening efforts for specific collection design [71] or to assess the natural product-likeness of de novo-generated natural product derivatives and collections of mimetics [20, 155]. The success of these studies is a remarkable testimony to the relevance of natural product fragments for the development of novel and impactful chemical probes and drug discovery hits [73].

5 Concluding Remarks and Outlook

As highlighted in this contribution, multiple impressive projects have relied on natural fragments to efficiently discover novel and selective starting points for drug discovery and chemical biology with great potential for further optimization [7, 9, 19, 20, 22]. The success of such endeavors stems, at least in part, from the potential to benefit from fragment-based approaches [17, 180] for compound discovery and design. At the same time, such natural fragments provide innovative, three-dimensional molecular frameworks [8, 17] that are chemically [4, 8] and computationally [2, 5, 7, 57, 107, 169] more accessible compared to their complex counterparts. Computational workflows have been implemented that support this process at all stages. Generating novel fragments [4, 9], their property analysis [8, 18, 57], their biological applications [4, 5, 21, 22], and compound collection designs [16, 20] can be supported through *in silico* approaches. It is noteworthy, however, that most of the computational tools discussed herein were not designed specifically with an application to natural products or their fragments in mind. Some early studies have indicated that algorithmic tools specifically tailored to process natural products might even further increase the performance and success rate of such pipelines [7, 9]. This indicates that further studies to design computational tools specifically for natural fragments, optimizing currently available workflows, and their intelligent application to drug discovery and chemical biology promise multiple avenues for impactful research and innovative molecular matter through coalescing data science and natural product research.

Acknowledgments Daniel Reker is a Swiss National Science Foundation Fellow (Grant P2EZP3_168827 and P300P2_177833). His work is partly supported by the MIT-IBM Watson AI Lab and the MIT SenseTime Alliance. Daniel Reker is grateful to his Ph.D. advisor Prof. Gisbert Schneider and his postdoctoral advisors Prof. Robert S. Langer and Prof. Giovanni Traverso for invaluable guidance and mentorship.

References

1. Ganesan A (2008) The impact of natural products upon modern drug discovery. *Curr Opin Chem Biol* 12:306
2. Barnes EC, Kumar R, Davis RA (2016) The use of isolated natural products as scaffolds for the generation of chemically diverse screening libraries for drug discovery. *Nat Prod Rep* 33:372
3. Harvey AL (2008) Natural products in drug discovery. *Drug Discov Today* 13:894
4. Over B, Wetzel S, Grütter C, Nakai Y, Renner S, Rauh D, Waldmann H (2013) Natural-product-derived fragments for fragment-based ligand discovery. *Nat Chem* 5:21
5. Rodrigues T, Reker D, Schneider P, Schneider G (2016) Counting on natural products for drug design. *Nat Chem* 8:531
6. Rollinger JM, Stuppner H, Langer T (2008) Virtual screening for the discovery of bioactive natural products. In: Petersen F, Amstutz R (eds) *Natural compounds as drugs*, vol 1. Birkhäuser, Basel, p 211
7. Reker D, Perna AM, Rodrigues T, Schneider P, Reutlinger M, Mönch B, Koeberle A, Lamers C, Gabler M, Steinmetz H, Müller R, Schubert-Zsilavec M, Werz O, Schneider G (2014) Revealing the macromolecular targets of complex natural products. *Nat Chem* 6:1072
8. Prescher H, Koch G, Schuhmann T, Ertl P, Bussenault A, Glick M, Dix I, Petersen F, Lizos DE (2017) Construction of a 3D-shaped, natural product like fragment library by fragmentation and diversification of natural products. *Bioorg Med Chem Lett* 25:921
9. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Caaulta M, Odermatt A, Ertl P, Waldmann H (2005) Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc Natl Acad Sci U S A* 102:17272
10. Dobson CM (2004) Chemical space and biology. *Nature* 432:824
11. Irwin JJ (2006) How good is your screening library? *Curr Opin Chem Biol* 10:352
12. Clardy J, Walsh C (2004) Lessons from natural molecules. *Nature* 432:829
13. Shoichet BK (2013) Nature's pieces. *Nat Chem* 5:9
14. Koeberle A, Werz O (2015) Multi-target approach for natural products in inflammation. *Drug Discov Today* 19:1871
15. Lewell XQ, Judd DB, Watson SP, Hann MM (1998) RECAP – retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci* 38:511
16. Wetzel S, Bon RS, Kumar K, Waldmann H (2011) Biology-oriented synthesis. *Angew Chem Int Ed* 50:10800
17. Vu H, Pedro L, Mak T, McCormick B, Rowley J, Liu M, Di Capua A, Williams-Noonan B, Pham NB, Rouwer R, Nguyen B, Andrew KT, Skinner-Adams T, Kim J, Hol WGJ, Hui R, Crowther GJ, Van Voorhis WC, Quinn RJ (2018) Fragment-based screening of a natural product library against 62 potential malaria drug targets employing native mass spectrometry. *ACS Infect Dis* 4:431
18. Pascolutti M, Campitelli M, Nguyen B, Pham N, Gorse A-D, Quinn RJ (2015) Capturing Nature's diversity. *PLoS One* 10:e0120942
19. Wetzel S, Klein K, Renner S, Rauh D, Oprea TI, Mutzel P, Waldmann H (2009) Interactive exploration of chemical space with Scaffold Hunter. *Nat Chem Biol* 5:581
20. Merk D, Grisoni F, Friedrich L, Gelzinyte E, Schneider G (2018) Computer-assisted discovery of retinoid X receptor modulating natural products and isofunctional mimetics. *J Med Chem* 61:5442
21. Rodrigues T, Reker D, Kunze J, Schneider P, Schneider G (2015) Revealing the macromolecular targets of fragment-like natural products. *Angew Chem Int Ed* 54:10516
22. Lanz J, Riedl R (2015) Merging allosteric and active site binding motifs: de novo generation of target selectivity and potency via natural-product-derived fragments. *ChemMedChem* 10:451

23. Rollinger JM, Hornick A, Langer T, Stuppner H, Prast H (2004) Acetylcholinesterase inhibitory activity of scopolin and scopoletin discovered by virtual screening of natural products. *J Med Chem* 47:6248
24. Hartenfeller M, Zettl H, Walter M, Rupp M, Reisen F, Proschak E, Weggen S, Stark H, Schneider G (2012) DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comp Biol* 8:e1002380
25. Reker D, Rodrigues T, Schneider P, Schneider G (2014) Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc Natl Acad Sci U S A* 111:4067
26. Chen Y, de Bruyn Kops C, Kirchmair J (2017) Data resources for the computer-guided discovery of bioactive natural products. *J Chem Inf Model* 57:2099
27. Natural Fragment Library|Pierre Fabre (2018) <https://www.pierre-fabre.com/en/natural-fragment-library>
28. BIONET Fragments from Nature (2019) <https://www.keyorganics.net/downloads-bionet-databases/>
29. Life chemicals – natural product-like fragment library (2019) <https://lifechemicals.com/screening-libraries/fragment-libraries#natural-lib>
30. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. *Nat Chem* 4:90
31. Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. *Nature* 432:855
32. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 46:3
33. Bemis GW, Murcko MA (1996) The properties of known drugs 1. Molecular frameworks. *J Med Chem* 39:2887
34. Landrum G (2012) RDKit: open-source cheminformatics. <https://www.rdkit.org>
35. Shang J, Hu B, Wang J, Zhu F, Kang Y, Li D, Sun H, Kong D-X, Hou T (2018) A cheminformatic insight into the differences between terrestrial and marine originated natural products. *J Chem Inf Model* 58:1182
36. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein-ligand docking using GOLD. *Proteins* 52:609
37. Congreve M, Carr R, Murray C, Jhoti H (2003) A “rule of three” for fragment-based lead discovery? *Drug Discov Today* 8:876
38. Todeschini R, Consonni V (2008) Handbook of molecular descriptors. Wiley-VCH, Weinheim
39. Labute P (2000) A widely applicable set of descriptors. *J Mol Graphics Model* 18:464
40. Allu TK, Oprea TI (2005) Rapid evaluation of synthetic and molecular complexity for in silico chemistry. *J Chem Inf Model* 45:1237
41. Ntie-Kang F, Onguéné PA, Scharfe M, Owono Owono LC, Megnassan E, Mbaze LM, Sippl W, Efange SMN (2014) ConMedNP: a natural product library from Central African medicinal plants for drug discovery. *RSC Adv* 4:409
42. Hall C & The Chapman & Hall/CRC Chemical Database Dictionary of Natural Products. Chapman and Hall/CRC, available at dnp.chemnetbase.com
43. Chen CY-C (2011) TCM database@Taiwan: the world’s largest traditional Chinese medicine database for drug screening in silico. *PLoS One* 6:e15939
44. Ertl P, Roggo S, Schuffenhauer A (2008) Natural product-likeness score and its application for prioritization of compound libraries. *J Chem Inf Model* 48:68
45. Jain AN (2004) Ligand-based structural hypotheses for virtual screening. *J Med Chem* 47:947
46. Klopman G (1984) Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J Am Chem Soc* 106:7315

47. Rosenkranz HS, Liu M, Cunningham A, Klopman G (1996) Application of structural concepts to evaluate the potential carcinogenicity of natural products. *SAR QSAR Environ Res* 5:79
48. Schreyer A, Blundell T (2009) CREDO: A protein-ligand interaction database for drug discovery. *Chem Biol Drug Des* 73:157
49. ChemicalComputingGroup (2011) Molecular operating environment (MOE) 2011.10. 1010 Sherbooke St. West, Suite #910, Montreal QC, Canada, H3A 2R7
50. Schneider G (2013) De novo molecular design. Wiley-VCH, Weinheim
51. Coley CW, Rogers L, Green WH, Jensen KF (2017) Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent Sci* 3:1237
52. Coley CW, Green WH, Jensen KF (2018) Machine learning in computer-aided synthesis planning. *Acc Chem Res* 51:1281
53. Langdon SR, Ertl P, Brown N (2010) Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Mol Inf* 29:366
54. Hu Y, Stumpfe D, Bajorath J (2016) Computational exploration of molecular scaffolds in medicinal chemistry. *J Med Chem* 59:4062
55. Schneider G, Schneider P, Renner S (2006) Scaffold-hopping: how far can you jump? *QSAR Comb Sci* 25:1162
56. Kontijevskis A (2017) Mapping of drug-like chemical universe with reduced complexity molecular frameworks. *J Chem Inf Model* 57:680
57. Grabowski K, Baringhaus K-H, Schneider G (2008) Scaffold diversity of natural products: inspiration for combinatorial library design. *Nat Prod Rep* 25:892
58. Lee M-L, Schneider G (2001) Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries. *J Comb Chem* 3:284
59. Boldi AM (2004) Libraries from natural product-like scaffolds. *Curr Opin Chem Biol* 8:281
60. Feher M, Schmidt JM (2002) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J Chem Inf Model* 43:218
61. Valli M, Dos Santos RN, Figueira LD, Nakajima CH, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2013) Development of a natural products database from the biodiversity of Brazil. *J Nat Prod* 76:439
62. Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, Mbah JA, Mbaze LM, Sippl W, Efang SMN (2013) AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PLoS One* 8:e78085
63. Dictionary of marine natural products. <http://dmnp.chemnetbase.com>
64. Kong D-X, Jiang Y-Y, Zhang H-Y (2010) Marine natural products as sources of novel scaffolds: achievement and concern. *Drug Discov Today* 15:884
65. Berthold MR, Cebon N, Dill F, Gabriel TR, Kötter T, Mainl T, Ohl P, Thiel K, Wiswedel B (2008) KNIME: The Konstanz information miner. *Data Anal Mach Learn Appl*:319
66. Henkel T, Brunne RM, Müller H, Reichel F (1999) Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angew Chem Int Ed* 38:643
67. Miyao T, Reker D, Schneider P, Funatsu K, Schneider G (2015) Chemography of natural product space. *Planta Med* 81:429
68. Oprea TI, Gottfries J (2001) Chemography: the art of navigating in chemical space. *J Comb Chem* 3:157
69. Reutlinger M, Schneider G (2012) Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J Mol Graphics Model* 34:108
70. Bajorath J, Peltason L, Wawer M, Guha R, Lajiness MS, Van Drie JH (2009) Navigating structure-activity landscapes. *Drug Discov Today* 14:698
71. Quinn RJ, Carroll AR, Pham NB, Baron P, Palframan ME, Suraweera L, Pierens GK, Muresan S (2008) Developing a drug-like natural product library. *J Nat Prod* 71:464
72. Charifson PS, Walters WP (2000) Filtering databases and chemical libraries. *Mol Divers* 5:185
73. Crane EA, Gademann K (2016) Capturing biological activity in natural product fragments by chemical synthesis. *Angew Chem Int Ed* 55:3882

74. Sadowski J, Gasteiger J (1993) From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem Rev* 93:2567
75. Maeda MH (2015) Current challenges in development of a database of three-dimensional chemical structures. *Front Bioeng Biotechnol* 3:66
76. Schwab CH (2010) Conformations and 3D pharmacophore searching. *Drug Discov Today Technol* 7:e245
77. Albady MA, Elokely KM, Wang B, Bowling JJ, Abdelwahab MF, Hossein MH, Doerksen RJ, Hamann MT (2013) Computationally assisted assignment of kahalalide Y configuration using an NMR-constrained conformational search. *J Nat Prod* 76:178
78. Rupp M, Bauer MR, Wilcken R, Lange A, Reutlinger M, Boeckler FM, Schneider G (2014) Machine learning estimates of natural product conformational energies. *PLoS Comput Biol* 10:e1003400
79. Rollinger JM, Schuster D, Danzl B, Schwaiger S, Markt P, Schmidtke M, Gertsch J, Raduner S, Wolberg G, Langer T, Stuppner H (2009) In silico target fishing for rationalized ligand discovery exemplified on constituents of *Ruta graveolens*. *Planta Med* 75:195
80. Atanasov AG, Wang JN, Gu SP, Bu J, Kramer M, Baumgartner L, Fakhrudin N, Ladurner A, Malainer C, Vuorinen A, Noha SM, Schwaiger S, Rollinger JM, Schuster D, Stuppner H, Dirsch VM, Heiss E (2013) Honokiol: a non-adipogenic PPAR γ agonist from Nature. *Biochim Biophys Acta, Gen Subj* 1830:4813
81. Sauer WHB, Schwarz MK (2003) Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J Chem Inf Model* 43:987
82. Olmedo DA, González-Medina M, Gupta MP, Medina-Franco JL (2017) Cheminformatic characterization of natural products from Panama. *Mol Divers* 21:779
83. Ertl P, Schuffenhauer A (2008) Cheminformatics analysis of natural products: lessons from Nature inspiring the design of new drugs. *Prog Drug Res* 66:217
84. Lankadurai BP, Nagato EG, Simpson MJ (2013) Environmental metabolomics: an emerging approach to study organism responses to environmental stressors. *Environ Rev* 21:180
85. Baker M (2011) Metabolomics: from small molecules to big ideas. *Nat Methods* 8:117
86. Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminform* 1:8
87. Grisoni F, Merk D, Consonni V, Hiss JA, Tagliabue SG, Todeschini R, Schneider G (2018) Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Commun Chem* 1:44
88. Wetzel S, Wilk W, Chammas S, Sperl B, Roth AG, Yektaoglu A, Renner S, Berg T, Arenz C, Giannis A, Oprea TI, Rauh D, Kaiser M, Waldmann H (2010) A scaffold-tree-merging strategy for prospective bioactivity annotation of γ -pyrones. *Angew Chem Int Ed* 49:3666
89. Böhm HJ, Flohr A, Stahl M (2004) Scaffold hopping. *Drug Discov Today Technol* 1:217
90. Schröder P, Förster T, Kleine S, Becker C, Richters A, Ziegler S, Rauh D, Kumar K, Waldmann H (2015) Neuritogenic militarinone-inspired 4-hydroxypyridones target the stress pathway kinase MAP4K4. *Angew Chem Int Ed* 54:12398
91. Rodrigues T (2017) Harnessing the potential of natural products in drug discovery from a cheminformatics vantage point. *Org Biomol Chem* 15:9275
92. Merk D, Grisoni F, Friedrich L, Schneider G (2018) Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Commun Chem* 1:68
93. Hopkins AL (2009) Drug discovery: predicting promiscuity. *Nature* 462:167
94. Hopkins AL, Groom CR, Alex A (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov Today* 9:430
95. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijter MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KLH, Edwards DD, Shoichet BK, Roth BL (2009) Predicting new molecular targets for known drugs. *Nature* 462:175
96. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* 53:2719

97. Reker D, Bernardes GJL, Rodrigues T (2019) Computational advances in combating colloidal aggregation in drug discovery. *Nat Chem* 11:402
98. Aldrich C, Bertozzi C, Georg GI, Kiessling L, Lindsley C, Liotta D, Merz KM Jr, Schepartz A, Wang S (2017) The ecstasy and agony of assay interference compounds. *ACS Cent Sci* 3:143
99. Bisson J, McAlpine JB, Friesen JB, Chen S-N, Graham J, Pauli GF (2016) Can invalid bioactives undermine natural product-based drug discovery? *J Med Chem* 59:1671
100. Huth JR, Mendoza R, Olejniczak ET, Johnson RW, Cothron DA, Liu Y, Lerner CG, Chen J, Hajduk PJ (2005) ALARM NMR: a rapid and robust experimental method to detect reactive false positives in biochemical screens. *J Am Chem Soc* 127:217
101. Baell JB (2016) Feeling Nature's PAINS: natural products, natural product drugs, and pan assay interference compounds (PAINS). *J Nat Prod* 79:616
102. Duan D, Doak AK, Nedyalkova L, Shoichet BK (2015) Colloidal aggregation and the in vitro activity of traditional Chinese medicines. *ACS Chem Biol* 10:978
103. Röhrig UF, Majjigapu SR, Vogel P, Zoete V, Michielin O (2015) Challenges in the discovery of indoleamine 2,3-dioxygenase 1 (IDO1) inhibitors. *J Med Chem* 58:9421
104. Glaser J, Holzgrabe U (2016) Focus on PAINS: false friends in the quest for selective anti-protozoal lead structures from Nature? *Med Chem Commun* 7:214
105. Dahlin JL, Nelson KM, Strasser JM, Baryte-Lovejoy D, Szweczyk MM, Organ S, Cuellar M, Singh G, Shrimp JH, Nguyen N, Meier JL, Arrowsmith CH, Brown PJ, Baell JB, Walters MA (2017) Assay interference and off-target liabilities of reported histone acetyltransferase inhibitors. *Nat Commun* 8:1527
106. Matlock MK, Hughes TB, Dahlin JL, Swamidass SJ (2018) Modeling small-molecule reactivity identifies promiscuous bioactive compounds. *J Chem Inf Model* 58:1483
107. Ekins S, Freundlich JS (2011) Validating new tuberculosis computational models with public whole cell screening aerobic activity datasets. *Pharm Res* 28:1859
108. Erlanson D (2013) Fragmenting natural products – sometimes PAINfully. <http://practicalfragments.blogspot.com/2013/02/fragmenting-natural-products-sometimes.html>. Accessed 18 Jun 2018
109. Irwin JJ, Duan D, Torosyan H, Doak AK, Ziebart KT, Sterling T, Tumanian G, Shoichet BK (2015) An aggregation advisor for ligand discovery. *J Med Chem* 58:7076
110. Pohjala L, Tammela P, Pohjala L, Tammela P (2012) Aggregating behavior of phenolic compounds – a source of false bioassay results? *Molecules* 17:10774
111. Rao H, Li Z, Li X, Ma X, Ung C, Li H, Liu X, Chen Y (2010) Identification of small molecule aggregators from large compound libraries by support vector machines. *J Comput Chem* 31:752
112. Feng BY, Shelat A, Dorman TN, Guy RK, Shoichet BK (2005) High-throughput assays for promiscuous inhibitors. *Nat Chem Biol* 1:146
113. Weaver S, Gleeson MP (2008) The importance of the domain of applicability in QSAR modeling. *J Mol Graphics Model* 26:1315
114. Dragos H, Gilles M, Alexandre V (2009) Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J Chem Inf Model* 49:1762
115. Ingólfsson HI, Thakur P, Herold KF, Hobart EA, Ramsey NB, Periole X, de Jong DH, Zwama M, Yilmaz D, Hall K, Marezky T, Hemmings HC, Blobel C, Marrink SJ, Koçer A (2014) Phytochemicals perturb membranes and promiscuously alter protein function. *ACS Chem Biol* 9:1788
116. Cristani M, D'Arrigo M, Mandalari G, Castelli F, Grazia Sarpietro M, Micieli D, Venuti V, Bisignano G, Saija A, Trombetta D (2007) Interaction of four monoterpenes contained in essential oils with model membranes: implications for their antibacterial activity. *J Agric Food Chem* 55:6300
117. Ossman T, Fabre G, Trouillas P (2016) Interaction of wine anthocyanin derivatives with lipid bilayer membranes. *Comput Theor Chem* 1077:80

118. Pillong M, Hiss JA, Schneider P, Lin B, Blatter M, Müller AT, Bachler S, Neuhaus CS, Ditttrich PS, Altmann KH, Wessler S, Schneider G (2017) Rational design of membrane-pore-forming peptides. *Small* 13:1701316
119. Lyu Y, Xiang N, Mondal J, Zhu X, Narsimhan G (2018) Characterization of interactions between curcumin and different types of lipid bilayers by molecular dynamics simulation. *J Phys Chem B* 122:2341
120. Capuzzi SJ, Muratov EN, Tropsha A (2017) Phantom PAINS: problems with the utility of alerts for pan-assay interference compounds. *J Chem Inf Model* 57:417
121. Jasial S, Hu Y, Bajorath J (2017) How frequently are pan-assay interference compounds active? Large-scale analysis of screening data reveals diverse activity profiles, low global hit frequency, and many consistently inactive compounds. *J Med Chem* 60:3879
122. Vidler LR, Watson IA, Margolis BJ, Cummins DJ, Brunavs M (2018) Investigating the behavior of published PAINS alerts using a pharmaceutical company data set. *ACS Med Chem Lett* 9:792
123. Owen SC, Doak AK, Wassam P, Shoichet MS, Shoichet BK (2012) Colloidal aggregation affects the efficacy of anticancer drugs in cell culture. *ACS Chem Biol* 7:1429
124. Bael JB, Nissink JWM (2018) Seven year itch: pan-assay interference compounds (PAINS) in 2017 – utility and limitations. *ACS Chem Biol* 13:36
125. Lagunin A, Filimonov D, Poroikov V (2010) Multi-targeted natural products evaluation based on biological activity prediction with PASS. *Curr Pharm Des* 16:1703
126. Sá MS, de Menezes MN, Krettli AU, Ribeiro IM, Tomassini TC, Ribeiro dos Santos R, de Azevedo WF, Soares MB (2011) Antimalarial activity of physalins B, D, F, and G. *J Nat Prod* 74:2269
127. Johnson MA, Maggiora GM (1990) Concepts and applications of molecular similarity. Wiley, New York
128. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50:742
129. MACCS Structural Keys (2005) – MDL Information Systems Inc
130. Nickel J, Gohlke B-O, Erehman J, Banerjee P, Rong WW, Goeade A, Dunkel M, Preissner R (2014) SuperPred: update on drug classification and target prediction. *Nucleic Acids Res* 42:W26
131. Antolín AA, Jalencas X, Yélamos J, Mestres J (2012) Identification of Pim kinases as novel targets for PJ34 with confounding effects in PARP biology. *ACS Chem Biol* 7:1962
132. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguez RM, Huang X-P, Norval S, Sassano MF, Shin AI, Webster LA, Simeons FRC, Stojanovski L, Prat A, Seidah NG, Constam DB, Bickerton GR, Read KD, Wetsel WC, Gilbert IH, Roth BL, Hopkins AL (2012) Automated design of ligands to polypharmacological profiles. *Nature* 492:215
133. Reutlinger M, Rodrigues T, Schneider P, Schneider G (2014) Combining on-chip synthesis of a focused combinatorial library with computational target prediction reveals imidazopyridine GPCR ligands. *Angew Chem Int Ed* 53:582
134. Schuster D, Spetea M, Music M, Rief S, Fink M, Kirchmair J, Schütz J, Wolber G, Langer T, Stuppner H, Schmidhammer H, Rollinger JM (2010) Morphinans and isoquinolines: Acetylcholinesterase inhibition, pharmacophore modeling, and interaction with opioid receptors. *Bioorg Med Chem Lett* 18:5071
135. Fakhruddin N, Ladurner A, Atanasov AG, Heiss EH, Baugartner L, Markt P, Schuster D, Ellmerer EP, Wolber G, Rollinger JM, Stuppner H, Dirsch VM (2010) Computer-aided discovery, validation, and mechanistic characterization of novel neolignan activators of peroxisome proliferator-activated receptor. *Mol Pharmacol* 77:559
136. Schneider G, Böhm H-J (2002) Virtual screening and fast automated docking methods. *Drug Discov Today* 7:64
137. Kitchen DB, Decomez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* 3:935
138. Lyne PD (2002) Structure-based virtual screening: an overview. *Drug Discov Today* 7:1047

139. Do Q-T, Lamy C, Renimel I, Sauvan N, Andre P, Himbert F, Morin-Allroy L, Bernard P (2007) Reverse pharmacognosy: identifying biological properties for plants by means of their molecule constituents: application to meranzin. *Planta Med* 73:1235
140. Bajorath J (2002) Affinity fingerprints – leading the way? *Drug Discov Today* 7:1035
141. Wassermann AM, Loukine E, Urban L, Whitebread S, Chen S, Hughes K, Guo H, Kutlina E, Fekete A, Klumpp M, Glick M (2014) A screening pattern recognition method finds new and divergent targets for drugs and natural products. *ACS Chem Biol* 9:1622
142. Reutlinger M, Koch CP, Reker D, Schneider TN, Rodrigues T, Schneider G (2013) Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for “orphan” molecules. *Mol Inf* 32:133
143. Schneider G, Neidhart W, Giller T, Schmid G (1999) “Scaffold-hopping” by topological pharmacophore search: a contribution to virtual screening. *Angew Chem Int Ed* 38:2894
144. Kohonen T (1990) The self-organizing map. *Proc IEEE* 78:1464
145. Schneider P, Tanrikulu Y, Schneider G (2009) Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing. *Curr Med Chem* 16:258
146. Digles D, Ecker GF (2011) Self-organizing maps for in silico screening and data visualization. *Mol Inf* 30:838
147. Reker D, Seet M, Pillong M, Koch CP, Schneider P, Witschel MC, Rottmann M, Freymond C, Brun R, Schweizer B, Illarionov B, Bacher A, Fischer M, Diederich F, Schneider G (2014) Deorphaning pyrrolopyrazines as potent multi-target antimalarial agents. *Angew Chem Int Ed* 53:7079
148. Schneider G, Reker D, Chen T, Hauenstein K, Schneider P, Altmann K-H (2016) Deorphaning the macromolecular targets of the natural anticancer compound dolicolide. *Angew Chem Int Ed* 55:12408
149. Rodrigues T, Werner M, Roth J, da Cruz EHG, Marques MC, Akkapeddi P, Lobo SA, Koeberle A, Corzana F, da Silva Junior EN, Werz O, Bernardes HJL (2018) Machine intelligence decrypts β -lapachone as an allosteric 5-lipoxygenase inhibitor. *Chem Sci* 9:6899
150. Schneider P, Schneider G (2017) A computational method for unveiling the target promiscuity of pharmacologically active compounds. *Angew Chem Int Ed* 56:11520
151. Joseph JA, Fisher DR, Cheng V, Rimando AM, Shukitt-Hale B (2008) Cellular and behavioral effects of stilbene resveratrol analogues: implications for reducing the deleterious effects of aging. *J Agric Food Chem* 56:10544
152. Rollinger JM, Steindl TM, Schuster D, Kirchmair J, Anrain K, Ellmerer EP, Langer T, Stuppner H, Wutzler P, Schmidtke M (2008) Structure-based virtual screening for the discovery of natural inhibitors for human rhinovirus coat protein. *J Med Chem* 51:842
153. Brand S, Roy S, Schröder P, Rathmer B, Roos J, Kapoor S, Patil S, Pommerenke C, Maier T, Janning P, Eberth S, Steinhilber D, Schade D, Schneider G, Kumar K, Ziegler S, Waldmann H (2018) Combined proteomic and in silico target identification reveal a role for 5-lipoxygenase in developmental signaling pathways. *Cell Chem Biol* 25:1095
154. Schmeller T, Sauerwein M, Sporer F, Wink M, Müller WE (1994) Binding of quinolizidine alkaloids to nicotinic and muscarinic acetylcholine receptors. *J Nat Prod* 57:1316
155. Foley DJ, Craven PGE, Collins PM, Doveston RG, Aimon A, Talon R, Churcher I, von Delft F, Marsden SP, Nelson A (2017) Synthesis and demonstration of the biological relevance of sp^3 -rich scaffolds distantly related to natural product frameworks. *Chem Eur J* 23:15227
156. Engels MFM, Thielemans T, Verbinnen D, Tollaenere JP, Verbeek R (2000) CerBeruS: a system supporting the sequential screening process. *J Chem Inf Comput Sci* 40:241
157. Renner S, Van Otterlo WAL, Dominguez Seoane M, Möcklinghoff S, Steinhilber D, Brunsveld L, Rauh D, Waldmann H (2009) Bioactivity-guided mapping and navigation of chemical space. *Nat Chem Biol* 5:585
158. Koch MA, Wittenberg L-O, Basu S, Jeyrai DA, Gouroulidou E, Reinecke K, Odermatt A, Waldmann H (2004) Compound library development guided by protein structure similarity clustering and natural product structure. *Proc Natl Acad Sci U S A* 101:16721

159. Schmid F, Jessen HJ, Burch P, Gademann K (2013) Truncated militarinone fragments identified by total chemical synthesis induce neurite outgrowth. *Med Chem Commun* 4:135
160. Jessen HJ, Schumacher A, Shaw T, Pfaltz A, Gademann K (2011) A unified approach for the stereoselective total synthesis of pyridone alkaloids and their neurotogenic activity. *Angew Chem Int Ed* 50:4222
161. Bon RS, Waldmann H (2010) Bioactivity-guided navigation of chemical space. *Acc Chem Res* 43:1103
162. Gunasekera SP, McCarthy PJ, Kelly-Borges M, Lobkovsky E, Clardy J (1996) Dysidiolide: a novel protein phosphatase inhibitor from the Caribbean sponge *Dysidea etheria* de Laubenfels. *J Am Chem Soc* 118:8759
163. Pascolutti M, Quinn RJ (2014) Natural products as lead structures: chemical transformations to create lead-like libraries. *Drug Discov Today* 19:215
164. Friedrich L, Rodrigues T, Neuhaus CS, Schneider P, Schneider G (2016) From complex natural products to simple synthetic mimetics by computational de novo design. *Angew Chem Int Ed* 55:6789
165. Rupp M, Schröter T, Steri R, Zettl H, Proschak E, Hansen K, Rau O, Schwarz O, Müller-Kuhrt L, Schubert-Zsilavecz M, Müller K-R, Schneider G (2010) From machine learning to natural product derivatives selectively activating transcription factor PPAR. *ChemMedChem* 5:191
166. Schneider P, Schneider G (2016) De novo design at the edge of chaos. *J Med Chem* 59:4077
167. Schneider G (2013) De novo design–hop(p)ing against hope. *Drug Discov Today Technol* 10:e453
168. Rodrigues T, Roudnicky F, Koch CP, Kudoh T, Reker D, Detmar M, Schneider G (2013) De novo design and optimization of aurora A kinase inhibitors. *Chem Sci* 4:1229
169. Perna AM, Rodrigues T, Schmidt TP, Böhm M, Stutz K, Reker D, Pfeiffer B, Altmann K-H, Backert S, Wessler S, Schneider G (2015) Fragment-based de novo design reveals a small-molecule inhibitor of *Helicobacter pylori* HtrA. *Angew Chem* 127:10382
170. Rodrigues T, Kudoh T, Roudnicky F, Lim YF, Lin Y-C, Koch CP, Seno M, Detmar M, Schneider G (2013) Steering target selectivity and potency by fragment-based de novo drug design. *Angew Chem Int Ed* 52:10006
171. Schneider G, Hartenfeller M, Reutlinger M, Tanrikulu Y, Proschak E, Schneider P (2009) Voyages to the (un)known: adaptive design of bioactive compounds. *Trends Biotechnol* 27:18
172. Spänkuch B, Keppner S, Lange L, Rodrigues T, Zettl H, Koch CP, Reutlinger M, Hartenfeller M, Schneider P, Schneider G (2013) Drugs by numbers: reaction-driven de novo design of potent and selective anticancer leads. *Angew Chem Int Ed* 52:4676
173. Rodrigues T, Reker D, Welin M, Caldera M, Brunner C, Gabernet G, Schneider P, Walse B, Schneider G (2015) De novo fragment design for drug discovery and chemical biology. *Angew Chem Int Ed* 54:15079
174. Gupta A, Müller AT, Huisman BJH, Fuchs JA, Schneider P, Schneider G (2018) Generative recurrent networks for de novo drug design. *Mol Inf* 37:1700111
175. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent Sci* 4:120
176. Blaschke T, Olivecrona M, Engkvist O, Bajorath J, Chen H (2018) Application of generative autoencoder in de novo molecular design. *Mol Inf* 37:1700123
177. Renner S, Schneider G (2006) Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* 1:181
178. Newman DJ, Cragg GM (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod* 75:311
179. Jayaseelan KV, Moreno P, Truszkowski A, Ertl P, Steinbeck C (2012) Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinformatics* 13:106
180. Murray CW, Rees DC (2009) The rise of fragment-based drug discovery. *Nat Chem* 1:187



Daniel Reker received a B.Sc. in Computer Science (2010) from the Technical University in Darmstadt, Germany, and a M.Sc. in Computational Biology and Bioinformatics from the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland (2012). He conducted his Ph.D. degree at the Institute of Pharmaceutical Sciences at ETH Zurich under the supervision Prof. Gisbert Schneider, where he developed active machine learning workflows for drug screening and ligand-based target prediction methods for de novo-designed small molecules and complex natural products. Since 2016, he has been a Swiss National Science Foundation (SNSF) postdoctoral fellow at the Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology (MIT), under the supervision of Prof. Robert Langer and Prof. Giovanni Traverso. His current research focuses on clinical and biochemical data analysis and coalesces natural products and machine learning for drug delivery research.

Open-Access Activity Prediction Tools for Natural Products. Case Study: hERG Blockers



Fabian Mayr, Christian Vieider, Veronika Temml, Hermann Stuppner, and Daniela Schuster

Contents

1	Introduction	178
1.1	General Background on hERG Physiology	178
1.2	The hERG Protein Structure	180
1.3	Types of hERG Modulators and Binding Sites	181
1.4	Computational hERG Assessment of Natural Products	181
2	Computational Target Prediction	182
2.1	Ligand-Based Target Prediction Methods	184
2.1.1	Two-Dimensional Ligand-Based Similarity Searching	184
2.1.2	Three-Dimensional Ligand-Based Similarity Searching	187
2.1.3	Machine-Learning Applications in Chemical Similarity Searching	191
2.1.4	Ligand-Based Pharmacophore Modeling	195
2.1.5	Generation of Protein Structures for Molecular Modeling	196
2.1.6	Molecular Docking	203
2.2	Evaluating Target Prediction Models	205
3	Publicly Available Target Prediction Tools Suitable for hERG	208
3.1	Similarity Ensemble Approach	208
3.2	SuperPred	209
3.3	SwissTargetPrediction	211
3.4	HitPick	212

F. Mayr

Institute of Pharmacy/Pharmacognosy, University of Innsbruck, Innsbruck, Austria

Institute of Pharmacy/Pharmaceutical Chemistry, University of Innsbruck, Innsbruck, Austria

e-mail: F.Mayr@uibk.ac.at

C. Vieider

Institute of Pharmacy/Pharmaceutical Chemistry, University of Innsbruck, Innsbruck, Austria

V. Temml · H. Stuppner

Institute of Pharmacy/Pharmacognosy, University of Innsbruck, Innsbruck, Austria

e-mail: Veronika.Temml@uibk.ac.at; Hermann.Stuppner@uibk.ac.at

D. Schuster (✉)

Institute of Pharmacy/Pharmaceutical Chemistry, University of Innsbruck, Innsbruck, Austria

Department of Pharmaceutical and Medicinal Chemistry, Institute of Pharmacy, Paracelsus Medical University Salzburg, Salzburg, Austria

e-mail: Daniela.Schuster@pmu.ac.at

© Springer Nature Switzerland AG 2019

A. D. Kinghorn, H. Falk, S. Gibbons, J. Kobayashi, Y. Asakawa, J.-K. Liu (eds.),

Progress in the Chemistry of Organic Natural Products, Vol. 110,

https://doi.org/10.1007/978-3-030-14632-0_6

177

3.5	admetSAR	213
3.6	PASSonline	214
3.7	Pred-hERG	215
3.8	VirtualToxLab™	215
4	Results and Discussion	216
4.1	Study Setup	216
4.2	Dataset Used in This Study	217
4.3	Data Curation and Pre-processing	218
4.4	Post-processing of Data	219
4.4.1	Performance with Natural Products	221
4.4.2	Performance with Synthetic Compounds	223
4.4.3	Overall Performance	225
5	Conclusion and Outlook	228
	References	229

1 Introduction

1.1 General Background on hERG Physiology

The human ether-a-go-go-related gene (hERG) encodes the α -subunit of voltage-gated potassium ion channels. The so-called hERG channel plays a key role in cardiac myocytes, where it controls the efflux of potassium ions and thereby coordinates a regular heartbeat. The ion efflux is necessary for the repolarization of the cardiac action potential [1]. A hERG channel blockage leads to a prolongation of cardiac repolarization, which is shown in electrocardiogram measurements as prolonged QT interval. Such blockage can eventually result in a dangerous ventricular tachyarrhythmia, called Torsade de Pointes, which is potentially life threatening [2]. Several marketed drugs such as the antihistamines astemizole (**9**) and terfenadine (**10**) and even substances used to prevent arrhythmia [such as quinidine (**6**) and dofetilide (**13**)] had to be withdrawn from the market by regulatory agencies, because they triggered ventricular arrhythmia and led, in some cases, to sudden death [3]. Figure 1 shows a selection of such withdrawn drugs in addition to several natural products known to interact with the hERG channel. A comprehensive list of withdrawn drugs can be found in the WITHDRAWN database (<http://cheminfo.charite.de/withdrawn/>) [4].

Due to this critical role, the hERG channel has emerged as a highly important antitarget in drug discovery. In 2005, harmonized preclinical and clinical guidelines were issued to ensure hERG channel-related risk assessment in drug development and to prevent torsadogenic drugs from reaching the market [5, 6]. Consequently, QT prolongation has become a major cause of attrition throughout the drug development pipeline [7].

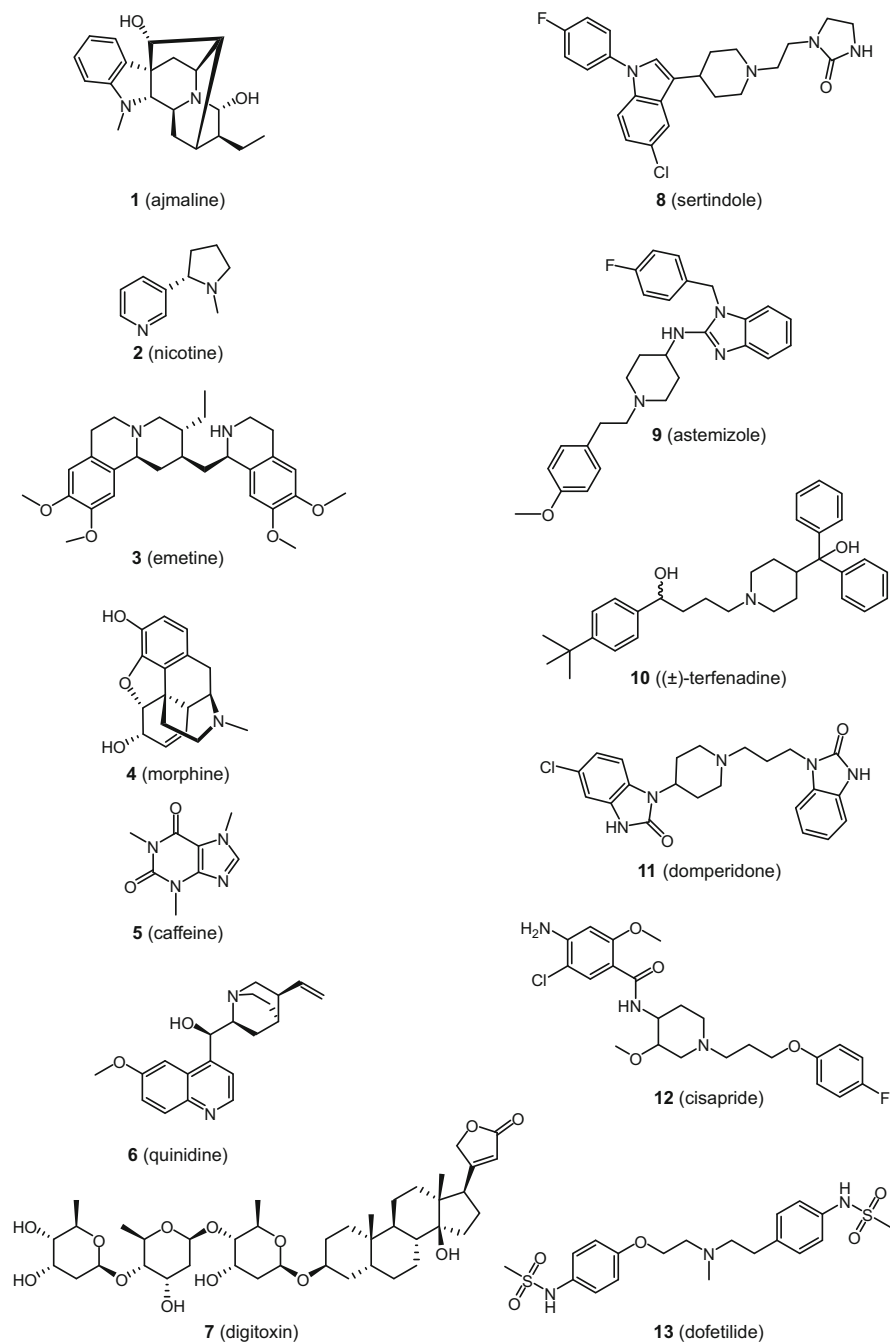


Fig. 1 Established natural products (1–7) that are well known for their hERG-blocking properties and drugs (8–13) that were withdrawn from the market due to hERG interaction

1.2 The hERG Protein Structure

Compounds interacting with hERG are structurally very diverse due to the high promiscuity of the hERG channel. The protein contains a large hydrophobic binding pocket, which undergoes conformational changes with the three states known for voltage-gated potassium channels (open, closed, and inactive) [8]. The channel is a tetramer, forming an ion channel pore. Each monomer consists of six transmembrane domains and amino- and carboxy-terminal cytoplasmic segments [1]. In the closed state, which is assumed at negative membrane potentials, the intracellular regions pucker up and close off the pore. Once depolarization occurs, the cell membrane allows for an opening of the channel for efflux of potassium ions. With further depolarization, the channel enters the inactive state, where the extracellular domains close the channel. These processes occur with a typical kinetic pattern of slow opening and closing and rapid voltage-dependent inactivation [1, 2, 9]. The architecture of the hERG channel is illustrated in Fig. 2. In 2017, Wang and MacKinnon reported a cryo-electron microscopy (cryo-EM) structure of the open hERG structure [10]. The three-dimensional models are available at the Protein Data Bank provided by the Research Collaboratory for Structural Bioinformatics

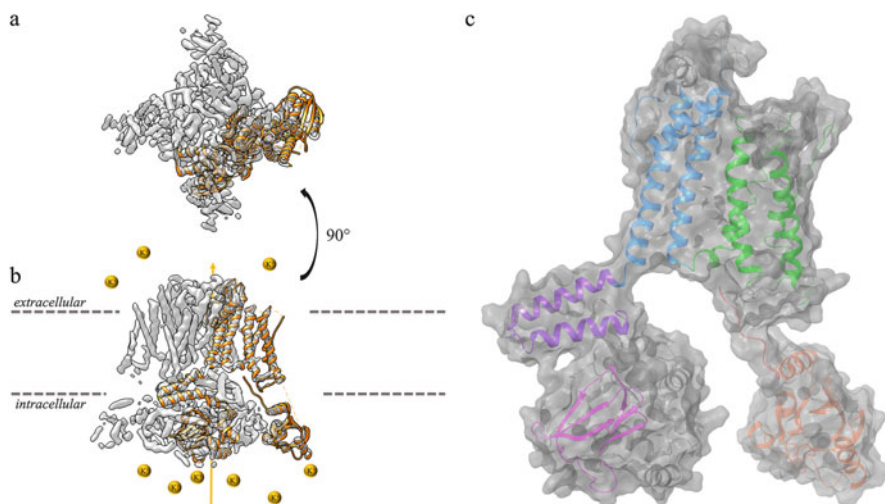


Fig. 2 Three-dimensional structure of hERG in the open conformation modified after Wang and MacKinnon [10]. **(a)** Top view (from extracellular) of hERG tetramer. Volume data as determined by cryo-EM are displayed as light gray shapes. Secondary structures were superimposed for one monomer (orange). **(b)** Side view of hERG tetramer embedded into the membrane, as obtained by 90° rotation of **(a)**. Schematically illustrated again are the secondary structures for one monomer (orange), the approximate membrane location (dotted line), the pore (yellow arrow), and potassium ions (yellow spheres). **(c)** Close-up of one hERG monomer in open conformation. The molecular surface is displayed as transparent, gray shape, with the respective secondary structures: Per Arnt Sirn (PAS)-domain (orange), voltage sensor (green), pore domain (blue), C-linker domain (purple), and cyclic nucleotide binding homology (CNBH) domain (pink)

(RCSB-PDB, <https://www.rcsb.org/>) under the four-letter codes “5VA1,” “5VA2,” and “5VA3.” The volume data as obtained by cryo-EM can be found in the EMDDataBank (EMD, <http://www.emdatabank.org/>) under the code “EMD-8650.” The most comprehensive database of experimentally solved three-dimensional structures of biomolecules is RCSB-PDB [11–14]. The EMD is the most comprehensive database for three-dimensional structures of biomolecules that were solved by cryo-EM [15]. Both databases are publicly available and are connected to one other.

The hERG channel binding modes have not been characterized comprehensively to this point. Due to the large binding pocket, the different conformational states, and the possibility of allosteric binding sites, it is likely that there are multiple ways to interfere with the channel.

1.3 Types of hERG Modulators and Binding Sites

There are different types of hERG modulators: blockers, which simply block the pore and lower electrolyte flux, and activators, which facilitate electrolyte flux at membrane potential almost high enough for channel activation [16]. For hERG channel blockers binding within the channel cavity, mutagenesis studies revealed the residues Tyr652 and Phe565 as crucial for ligand binding. For some blockers, also mutation of Thr623 and Ser624 led to reduced activity [8, 17]. More recently, Saxena et al. also identified Phe557 as a key residue for hERG blockage, which also plays a role for hERG activators [18]. All of the abovementioned amino acid residues are located either directly at the pore-forming helices or at the nearby located voltage sensor (Fig. 2c).

The hERG activators are divided into two classes, depending on how they enhance the hERG current. Type 1 hampers channel deactivation and reduces inactivation, while type 2 activators only reduce channel inactivation. A mutagenesis experiment on the type 1 hERG channel activator RPR260243 showed that mutations of the four residues Leu553, Phe557, Asp658, and Val659 to alanine abolished its activity [19]. For the type 2 hERG channel activator PD118057, mutations of Leu646 or Phe619 abolished activity [20]. Furthermore, an allosteric mechanism of action has been proposed for several compounds [21].

To sum up, it is probable that there are several hERG binding sites distributed across the ligand accessible regions of the tetramer. This circumstance complicates structure-based modeling efforts on the hERG channel.

1.4 Computational hERG Assessment of Natural Products

Methods to assess hERG-related cardiotoxicity in novel chemical structures with low cost are of high interest. This is, however, a highly complex field, due to the multiple ways to interfere with the hERG channel with different binding modes and

modes of action [22, 23]. Furthermore, channel promiscuity and interactions with other ion channels may prove problematic [1].

Experimentally, hERG activity is evaluated predominantly in a patch clamp assay, which is generally expensive and time-consuming. Attempts at using automated patch clamp systems exist but currently do not reach the same quality standards as the conventional experimental setups [24].

A wide array of *in silico* methods has been developed to predict hERG activity, as reviewed by Villoutreix and Taboureau [25]. However, *in silico* methods suffer from the poor understanding of binding modes on hERG. This realization, together with the relatively late availability of a three-dimensional structural model of human hERG, has forced *in silico* experiments to focus on ligand-based approaches.

While novel drug candidates are examined routinely for hERG-channel-related cardiotoxicity, many herbal remedies and even commonly consumed plants have not been tested for their hERG-related properties. As natural products play a vital role in the healthcare system of many developing countries and also constitute a large share of the drugs used overall, often accessible without prescription, it is crucial to examine them for hERG channel-related off-target effects.

Computational methods have been widely applied, and partly also successfully, to predict hERG blocking activity [25–27]. However, natural products often pose a special challenge for computational models because their physicochemical properties differ from what we typically see as “drug-like” molecules [28]. They generally possess a larger molecular mass, contain fewer rotatable bonds, more stereocenters, and more oxygens, to name just a few aspects. A comprehensive overview on the differences between synthetic drugs and natural products can be found in a review by Feher and Schmidt [29].

2 Computational Target Prediction

The central domain of the drug discovery scientist is the exploration of novel bioactivities. The bioactivity itself is typically a product of a compound’s interaction with a macromolecular target, mostly a protein [30]. Any bioactivity, next to its qualitative nature (compound A interacts with protein 1), can also be described using quantitative metrics (EC_{50} , IC_{50} , K_i , etc.). These data currently are readily available in public databases like ChEMBL [31–33] or PubChem [34, 35] and can be mined on a large scale due to recent advances in computer power and big data processing. The ever-growing content of such databases constitutes nothing less than the foundation of any cheminformatic approach used in current drug discovery. It allows for a systematic search for new bioactivities according to the similarity principle: similar compounds exert similar bioactivities [36, 37]—which is perhaps the most important hallmark of medicinal chemistry. The search for novel bioactivities can be approached from two directions: screening for new ligands of a molecular target or screening for new molecular targets of a ligand. The first of these applies to a typical virtual screening (VS), independent of the actual technique used (e.g.,

molecular docking, pharmacophore-based VS, 2D similarity searches, etc.). The second strategy allows for the identification of molecular targets for ligands and is therefore often referred to as target prediction, target fishing, activity profiling, or inverse virtual screening (iVS). This is undisputedly a younger but still an established approach. Nevertheless, target prediction has gained steeply in popularity and already plays a crucial role in the toolkit of a drug discovery scientist today. The main fields of application to target prediction are diverse: first, the de-orphanization of compounds with unknown interaction partners. Such compounds are often produced from phenotypic high-throughput screenings [38, 39], where cell-based assays are deployed, for which readout cannot be associated with one specific molecular target (e.g., pro-apoptotic or antimicrobial). In these cases, a so-called target deconvolution has to be conducted, which is time-consuming and challenging [40]. Even though several target deconvolution strategies are available and have been successful in the past, *in silico* target prediction can support a search for the respective molecular target in an economic manner [41]. Second, drug repositioning, or often referred to as “drug repurposing” in the literature, is yet another method in drug discovery to take advantage from *in silico* target prediction. Drug repositioning is aimed at establishing new indications for approved drugs. These drugs are used as input for the *in silico* target prediction, and the respective predictions are validated subsequently. The steep increase in the popularity of drug repositioning can be attributed primarily to the increasing pressure on research and development over the last decade. The great advantage of a drug repositioning campaign compared to a conventional drug discovery campaign is the considerably smaller effort in research and development incurred. Hence, pharmacokinetic studies, toxicity data, and expensive clinical data can be reutilized from previously conducted studies. The same is true for the development of an appropriate dosage form, production process, and the respective facilities. These circumstances save time and resources and ultimately result in shortened time and drug development costs [42]. Third, in the early stages of drug discovery, target prediction plays an important role in *in silico* ADMET prediction. The aim of this is to identify potential antitargets for the structure in question. Antitargets are druggable macromolecular structures that are mostly associated with side effects, like, e.g., various cytochrome P450 enzymes (CYP), P-glycoprotein, or hERG. Binding to such antitargets has to be avoided and is therefore determined at an early stage in the drug discovery process. Modern target prediction tools can facilitate the identification of such unwanted molecular targets of a lead compound, even before any examination in a wet laboratory. In this chapter, the focus is based on the third field of application: an *in silico* target prediction on hERG for natural products [43–46]. Fourth, especially in natural products research, target prediction is often applied to deconvolute the bioactivity of extracts. Extracts, or sub-fractions of extracts, are after all complex mixtures of a myriad of constituents. Target prediction can be used to address the bioactivity to certain natural products abundant in a bioactive extract.

Regardless of what the goal of the target prediction is, the same techniques are applied. They are subdivided into ligand-based and structure-based methods. This principle can be found, for example, in “reverse pharmacognosy.”

2.1 *Ligand-Based Target Prediction Methods*

Ligand-based methods are a widely used concept in cheminformatics today and are defined as models that are derived from the various available ligands for a specific target activity. The fundamental difference between structure-based models and ligand-based models is evident, as follows: while structure-based models are derived from protein-ligand complexes of the target of interest, ligand-based models are derived from at least one known ligand/s of this same target. Accordingly, the first step of any VS campaign is typically to check whether three-dimensional structures in adequate quality are available. Ligand-based models are therefore often seen incorrectly as the stopgaps in VS, if no three-dimensional structure can be found. However, and to the contrary, ligand-based models should be seen as a potent approach that brings into play a myriad of molecular targets, which three-dimensional structures have not been solved previously. For example, while at the beginning of 2019, 4563 distinct drug targets were listed in DrugBank (<https://www.drugbank.ca/>) [47], 1215 of those protein structures have been solved and are represented by at least 1 structure deposited in the RCSB-PDB. DrugBank is a comprehensive database of drugs, drug targets, and diseases of drugs that are either marketed or have been included in a clinical trial [48–52]. Both databases are publicly available. Even though these numbers may not represent the undisputed truth (it is simply not possible to categorize proteins into only drug targets and non-targets), they seem to represent the day-to-day subjective experience of a molecular modeler quite well: approximately 75% of the protein targets of interest for a drug discovery scientist will lack three-dimensional structural information. However, neither ligand-based approaches nor structure-based approaches can guarantee success. The truth is both approaches bear some inherent advantages as well as disadvantages. Ultimately, it has to be decided case-by-case which approach is the most promising for the specific project.

2.1.1 **Two-Dimensional Ligand-Based Similarity Searching**

Ligand-based approaches can be subdivided into two major classes: two-dimensional (2D) and three-dimensional (3D) approaches. Two-dimensional ligand-based models compare the similarity between two molecules on a 2D level, without considering conformations or spatial arrangement. Two-dimensional approaches solely focus on atom types and their connectivity using so-called molecular fingerprints. Molecular fingerprints are the bit string representation of molecular structures. They can be seen as lists of predefined features, whereas “1” indicates the presence and “0” indicates the absence of that feature in the molecule represented. The length of the fingerprints or the count of bits corresponds to the number of such features being listed. There are several types of fingerprints that can be computed. Cereto-Massague et al. grouped them into substructure keys-based fingerprints, topological or path-based fingerprints, and circular fingerprints [53]. Substructure

keys-based fingerprints are bit strings that are used to browse the query molecule for predefined substructures or features. The type of substructures and the order in the bit string are defined by the respective fingerprint. A popular example of this group would be the two MACCS fingerprints with different lengths [54]. Topological or path-based fingerprints browse for fragments of the query molecule in a typically linear pattern up to a predefined distance. Length, order, and type of fragment are defined by the fingerprint. An example for a topological fingerprint is the prominent Daylight fingerprint [55] utilized by the early Similarity Ensemble Approach (SEA) or the FP2 fingerprint used in SwissTargetPrediction. Finally, circular fingerprints browse for molecular fragments that make up the query molecule. In contrast to topological fingerprints, circular fingerprints work with circular concentric fragments of a molecule, starting from one atom like indicated schematically in Fig. 3. The number of such concentric circles used is specific to the respective fingerprint, usually specified in its name (e.g., the number “two” in FCFP2). A particularly popular class of fingerprints are the extended connectivity fingerprints (ECFPs). For example, SuperPred and admetSAR use the ECFP4 and the ECFP8 fingerprint, respectively, for 2D similarity comparison. HitPick makes use of the so-called functional class fingerprints (FCFP), a variant of ECFP. While ECFP aim to represent accurately the atomic environment of the center atom, FCFP describe the abstracted functional class environment of the center atom. Functional class fingerprints are often described as incorporating a pharmacophore-like idea into fingerprints, which are 2D representation of molecules [56]. PASSonline, on the other hand, uses the so-called multilevel neighborhood of atoms (MNA) to assess the 2D similarity of molecule pairs—yet another circular fingerprint [57].

The principal reason to use molecular fingerprints in VS is to compare the similarities of two molecules. To do so, the distance between the two bit strings representing the two molecules has to be calculated. The distance, a positive numerical value, can be interconverted to a similarity metric, a value between zero and one, using Eq. (1).

$$\textit{Similarity} = \frac{1}{1 + \textit{distance}} \quad (1)$$

Equation 1 Conversion from distance metric to a similarity metric

Even though several distance metrics are available that can be translated to a similarity metric, the most popular similarity coefficient in cheminformatics applications is the Tanimoto or Jaccard coefficient (T_c) [58]. The T_c is calculated according to Eq. (2). There are several similarity metrics available for such comparisons, the most popular being the Dice coefficient, the Cosine similarity, the Russel-RAO coefficient, and the Forbes coefficient (reviewed in [53]). In 2015, Bajusz et al. compared several of these similarity coefficients in the context of 2D similarity comparisons of molecular structures. The authors found T_c to be the most representative of chemical similarity comparisons [59].

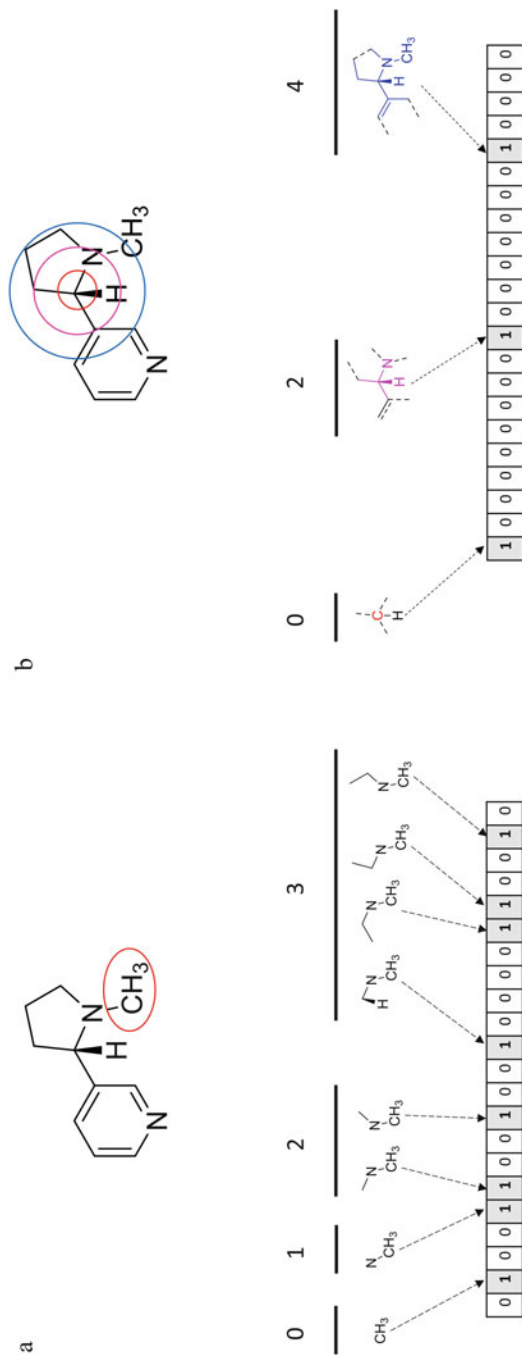


Fig. 3 Explanatory scheme of the generation of two distinct fingerprint types from the hERG blocker nicotine (2). Topological fingerprints are shown on the left side (b) and circular fingerprints on the right side (b), with subsequent translation into the respective bit strings

$$T_c = \frac{AB}{A + B - AB} \quad (2)$$

Equation 2 Calculation of Tanimoto coefficient (T_c) using molecular fingerprints. A indicates the amount of bits set to “1” in the first bit string, B indicates the amount of bits set to “1” in the second bit string, and AB indicates the amount of bits set to “1” in both bit strings

T_c s range from zero to one, while zero indicates no 2D similarity or, in other words, the highest possible dissimilarity, and one indicates identical molecules. For a more detailed explanation of molecular fingerprints, the interested reader may refer to [53, 60].

Molecular fingerprints are currently the state of the art to bring 2D molecular structures into a format, which is readable for a computer. Fingerprints have in common that both their generation and their comparison are “computationally cheap,” meaning that both processes are rapidly computed—regardless of which fingerprint is used. This fact constitutes the most important advantage of 2D similarity calculations: they are fast and easy to use, allowing for high-throughput rates at low computational costs. The biggest limitation of 2D similarity comparisons so far lies in the very nature of molecular fingerprints: the molecular similarity coefficients produced are strictly limited to the 2D representation of the compounds. However, molecules are three-dimensional entities, and therefore molecular fingerprints ignore an important property of molecules in that they may not accurately predict the activities of chiral compounds.

2.1.2 Three-Dimensional Ligand-Based Similarity Searching

Three-dimensional ligand-based models compare the three-dimensional “shape” of molecules to one another. This principle is, in contrast to 2D ligand-based models, based on the theory of molecular complementarity (ligand–target), rather than molecular similarity (ligand–ligand of target). A ligand can only bind to its target, if their shapes complement each other to a certain degree, and allow the ligand to fit perfectly into the binding pocket. Both of these principles are legitimate and belong to the same umbrella paradigm of today’s understanding of ligand–target interaction, namely, the key-lock principle, first hypothesized by Emil Fischer in 1894 [61]. Three-dimensional ligand-based models therefore aim to compare the molecular shapes of a target’s known ligands with a set of screening compounds. This is of particular interest in regard to a phenomenon addressed as “scaffold hopping.” Scaffold hopping is the ability of a method to identify chemically diverse scaffolds that bind to the same target. The similarity of the compounds shows in the third dimension, in that molecules appearing dissimilar in a 2D representation can be quite similar if looked upon in a three-dimensional representation. Three-dimensional ligand-based models were designed to account for exactly this circumstance and, in fact, are less susceptible to fail because of scaffold hopping. The algorithms behind the vast majority of three-dimensional ligand-based models belong to only two classes: Gaussian shape methods and approximate shape methods. Three-dimensional ligand-based models, especially

Gaussian shape methods, are generally considered a lot more exhaustive, due to their expansive calculations. Approximate shape methods were developed about 5 years later, aiming at reducing calculation times.

Gaussian shape methods derive their name from the spherically symmetric functions that are used to describe the atoms of the respective molecules. The overlap between the spheres of two molecules is then calculated and used as a quantitative estimate of the shape similarity. This approach is hampered by one important issue: to be able to compare two shapes that potentially are made up by different atoms and/or different amounts of atoms, these molecules need to be superimposed earlier. This can be achieved either by an algorithm that fits one molecular shape into another over iterative optimization cycles [62] or by a translation of all atoms into a common coordinate system, followed by the superimposition of the centers of masses and the alignment of the principal axes of inertia [63]. This process is illustrated schematically in Fig. 4. In both variants, various orientations of both molecules are sampled, because the two methods suffer from a considerable bias caused by the starting orientations. Neglecting this important issue can lead to a missed match between two molecular shapes due to different orientations.

Finally, the shape overlap $O_{A,B}$ of two correctly superimposed molecular shapes can be computed (for a detailed mathematical description, see [64]). $O_{A,B}$ then allows for the calculation of distance metrics (Euclidean distance, Manhattan distance) or

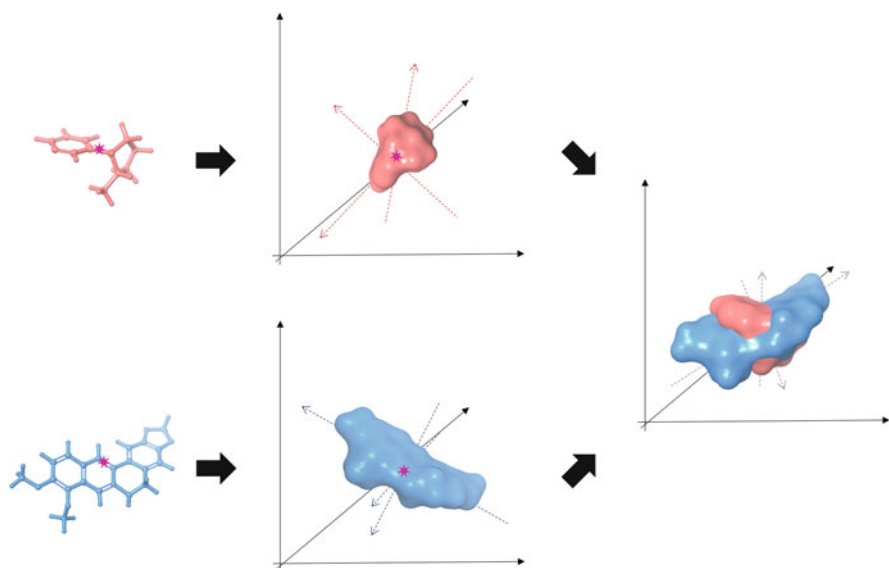


Fig. 4 Gaussian shape method for three-dimensional shape comparison of the hERG blockers nicotine (**2**) in light red, and dihydroberberine in light blue, schematically illustrated. A pink asterisk marks the centers of masses in both molecules; the three dotted arrows mark the computed principal moments of inertia

similarity metrics like T_c [see Eq. (3)], as shown before for two-dimensional ligand-based models.

$$T_c = \frac{O_{A,B}}{O_{A,A} + O_{B,B} + O_{A,B}} \quad (3)$$

Equation 3 Calculation of Tanimoto coefficient (T_c) using molecular shapes. $O_{A,B}$ is the shape overlap of molecules A , and molecule B , $O_{A,A}$ is the shape overlap of molecule A with itself, and $O_{B,B}$ is the shape overlap of molecule B with itself

The T_c obtained from such three-dimensional ligand-based models can be interpreted identical to T_{cS} obtained from 2D ligand-based models: a T_c of zero indicates no shape similarity, while a T_c of one indicates a perfectly congruent shape overlap. The first commercial application of a three-dimensional ligand-based model based on a Gaussian shape method was the Rapid Overlay of Chemical Structures (ROCS) [64, 65]. In terms of publicly available target prediction tools, SuperPred makes use of an algorithm based on a Gaussian shape method.

More recent advancements in three-dimensional ligand-based models are the so-called approximate shape methods. These algorithms are designed in a completely different manner, with the objective of accelerating the calculation time compared to the much slower Gaussian shape methods. To do so, the laborious steps of pre-aligning and orienting molecules and computing of overall shapes can be avoided. Instead, the molecular shape is “approximated” and described through different variables. Ballester et al. introduced the Ultra Shape Recognition (USR) to describe molecular shapes [66, 67]. First, four molecular locations are defined: the centroid of the molecule (ctd), the closest atom to ctd (cst), the farthest atom from ctd (fct), and the farthest atom from fct (ftf). Then, Euclidean distances are computed from all atoms to the four respective molecular locations. Each of the resulting four distances is described with three moments μ , namely, its average μ_1 , its standard deviation μ_2 , and its cubic root of the skewness μ_3 . The molecular shape \vec{M}_B is ultimately made up by 12 variables. This process is illustrated schematically in Fig. 5. Using this method, USR circumvents time-consuming calculation of absolute atom coordinates using relative coordinates. Moreover, the 12 variables are computed rapidly and are easily stored. This allows also for high throughput in an acceptable amount of time.

To compare how similar the shapes of two molecules are, the similarity metric, $S_{A,B}$, can be computed as indicated in Eq. (4). The result ranges again from zero to one, and the interpretation is consistent with T_{cS} .

$$S_{A,B} = \left(1 + \frac{1}{12} \sum_{k=1}^{12} \left| \vec{M}_A - \vec{M}_B \right| \right)^{-1} \quad (4)$$

Equation 4 Similarity score obtained from approximate shape methods. \vec{M}_A describes the molecular shape of molecule A and \vec{M}_B describes the molecular shape of molecule B , both containing 12 variables each

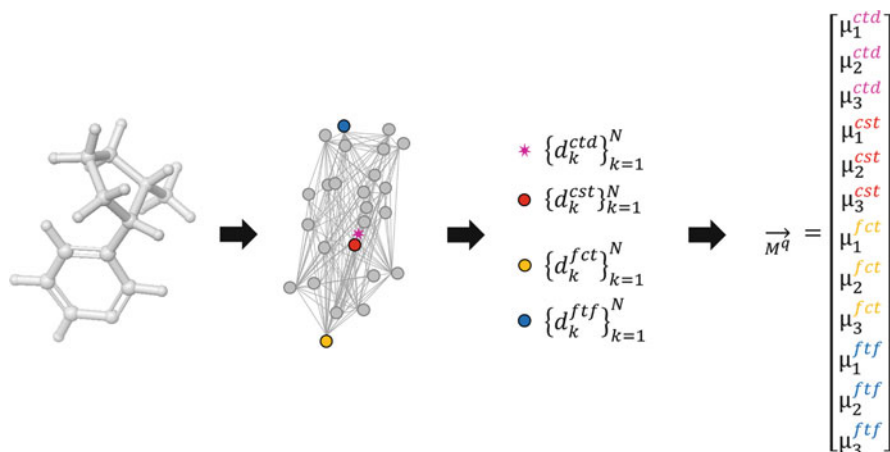


Fig. 5 Working principle of approximate shape methods schematically illustrated. For the example molecule nicotine (**2**), the centroid is computed (ctd), as well as cst, fct, and ftf. For each of these four molecular positions, 3 moments μ are computed, resulting in a 12-variable vector (\vec{M}_B)

Three-dimensional similarity measures are in general more sophisticated than 2D methods. Initially, they were even thought to have superior prediction power over 2D methods, for the simple reason that molecules are three-dimensional, rather than two-dimensional. Therefore, a three-dimensional comparison should perform better in, e.g., retrieving active compounds from a random screening list, and thus balance the higher computational cost. Unfortunately, this is not the case. In some cases, three-dimensional methods are even outperformed by two-dimensional methods [44]. This phenomenon can be attributed mainly to the fact that the shape of an active molecule alone does not explain the bioactivity. Important concepts like chemical functionalities, electronic properties of the molecule's surface, or chirality are neglected. These drawbacks were eliminated stepwise, especially within the approximate shape methods, e.g., by accounting for chirality as in Chiral Shape Recognition (CSR) [68] or adding two further dimensions like in ElectroShape Recognition (UFSRAT) [70] or Ultra-Fast Shape Recognition with CREDO Atom Types (USRCAT) [71]. Within the Gaussian shape methods, the same development took place. Noteworthy are the implementation of the hydrogen-bond propensity in ROCS and the development of SHApe-FeaTure Similarity (SHAFTS): a combination of shape similarity and pharmacophoric features. As a rule of thumb for the medicinal chemist, three-dimensional ligand-based methods are said to perform better for low structural similarity, because an ability to detect scaffold hopping comes into play. Within higher structural similarity, two-dimensional ligand-based methods usually perform better or at least equal [72]. For example, the three-dimensional shape comparison implemented in SwissTargetPrediction uses a USR-based algorithm.

2.1.3 Machine-Learning Applications in Chemical Similarity Searching

Artificial intelligence is currently one of the most dominant technologies in our daily lives. As artificial intelligence steadily advanced over the last decade, it has also been introduced into drug discovery. This has led to a myriad of applications and tools using machine learning over the last couple of years, which are discussed elsewhere [73–75]. The field of target prediction also has been subjected to experiments with such algorithms, especially in its subcategory of ligand-based target prediction. When discussing machine learning, usually, the respective algorithms are meant. These machine-learning algorithms (MLAs) use statistical tools that are able to “learn” from a given training set, without being explicitly coded. This makes MLAs particularly versatile and applicable to numerous problems, one being ligand-based target prediction. Machine learning can be subdivided into two main groups: supervised learning and unsupervised learning. In supervised learning, the algorithm “knows the correct answer,” meaning that it is trained with a number of exemplary input-output pairs. In unsupervised learning, the algorithm is not given the correct answer, because an absolutely correct answer may not exist. Instead, the algorithm detects commonalities and patterns in the test set, and, e.g., groups them accordingly (such as clustering of chemical structures). Supervised learning again consists of two subgroups, namely, regression models and classification models. While regression models are used typically to predict continuous numerical variables (e.g., demand for product “X” during holidays or relative humidity for a weather forecast), classification models aim to predict discrete classes (e.g., “tomorrow will be rainy” and “tomorrow won’t be rainy” or “inhibiting cyclooxygenase-2” and “not inhibiting cyclooxygenase-2”). As might be apparent to the attentive reader, the kind of MLA to employ depends greatly on how the question is posed, so there are problems that can be solved by both classification and regression models. In other words, by reformulating the question, another class of MLAs becomes usable. Despite this great variety and interchangeability of MLAs, almost exclusively classification models are used for VS as well as for target prediction. Exceptions are quantitative structure-activity relationship (QSAR) models, which use regression models to predict, e.g., activity values. The development of a machine-learning model is performed in three phases, as illustrated in Fig. 6a. First, the data have to be prepared. In the case of target prediction, a training set of molecules with a given class, typically “active” or “inactive” on the target of interest has to be gathered. This step is particularly crucial, because the quality and the nature of the dataset ultimately will determine the behavior of the model. Second, the MLA can be trained using the just-prepared training set. During this step, the algorithm will adjust its prediction to what it can “learn” from the training set. This means that the algorithm looks for features that discriminate well between the two classes “active” or “inactive.” For example, all active molecules bear a carbonyl group in position C-2, or most of the inactive molecules are substituted in position C-7. Third, based on the knowledge obtained during the training phase, the algorithm optimizes a model that can finally be deployed. The model can then be used for predicting the bioactivity of

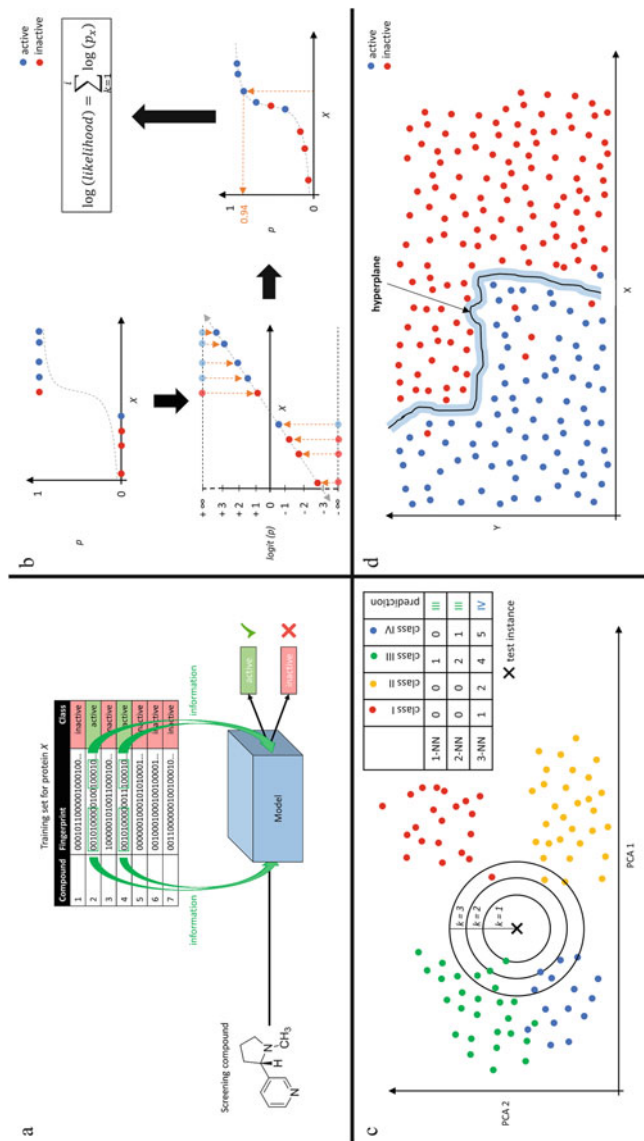


Fig. 6 Popular machine-learning algorithms (MLAs) for classification. **(a)** General working principle of MLAs: Patterns are recognized, which discriminate between two classes in a training set that are to be predicted, and deduced into a model. The resulting model can be applied to new instances as with the natural product hERG blocker nicotine (**2**). Compound **2** is converted to the same bit string as the training molecules, and assigned to one of the two classes “active” or “inactive,” based on what it has learned from the training set. **(b)** Curve fitting in a logistic regression. Red dots indicate “inactive” and blue dots indicate “active” instances. The logistic function is altered until the log(likelihood) reaches a minimum. **(c)** k-Nearest Neighbors (k-NN). Test instance is projected to the training data and the distances calculated. 1-NN and 2-NN both predict class III, while 3-NN predicts class IV. **(d)** Support vector machines (SVM). A hyperplane is generated inside the training data that separates the two classes. Here, a nonlinear kernel was used, since linear separation was not possible. Support vector machines try to maximize the distance between the hyperplane and the closest training instances (indicated in blue shade around the hyperplane)

new molecules that were not present in the training set. Here, the algorithm applies the knowledge it has obtained during the training phase on new, yet unclassified molecules. Some of the most popular MLAs, namely, naïve Bayesian classifiers, support vector machines (SVMs), logistic regressions, and *k*-nearest neighbors (*k*-NN), are mentioned in this chapter. Therefore, such MLAs require more detailed explanation.

Naïve Bayesian classifiers rely on the Bayes theorem, which aims to calculate the probability of an event *A* under a certain condition *B*. This so-called a-posteriori-probability $P(A|B)$ can be calculated by multiplying the probability of the event *B* happening under the condition *A* $P(B|A)$ with the a-priori-probability of *A*, $P(A)$, all divided by the a-priori-probability of *B*, $P(B)$, as shown in Eq. (5).

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)} \quad (5)$$

Equation 5 Bayes theorem to calculate the a-posteriori-probability of event *A* happening under the condition of *B*

Such Bayesian models are typically referred to as “naïve,” due to their simplified assumption of strong independence between the various features. This assumption simplifies the calculation significantly, even though feature independence of this extent rarely occurs in real-world examples. An advantage is that due to assumed feature independence, no normalization is necessary. Nevertheless, naïve Bayesian classifiers have proven successful in VS and related fields, but also, e.g., as spam filters or document classifiers. Another term frequently used in association with naïve Bayes is “Laplacian modification.” The Laplacian modification, also known as Laplacian smoothening, introduces a so-called pseudo count ensuring that no probability is ever set to zero. Without the Laplacian modification, naïve Bayesian classifiers would set the probability of belonging to a class that was never associated to a certain feature value in the training set always to zero. Such zero probabilities disrupt further calculations and unnaturally lower the overall probability. Applications of such Bayesian classifiers will be presented later in HitPick (Sect. 3.4), admetSAR (Sect. 3.5), and PASSonline (Sect. 3.6).

Logistic regression is yet another MLA, typically used to classify test instances into two discrete classes (e.g., “active” and “inactive”) from one continuous attribute *x* (e.g., similarity score). The logistic regression belongs to supervised learning. Typically, the continuous variable is displayed on the *x*-axis, while the *y*-axis displays the probabilities from zero to one (see Fig. 6b). Despite the *y*-axis covering a continuous range of values, the final output of a logistic regression is binominal, with probability values of either zero or one. The logistic regression is, seen from its basic idea, close to the common linear regression, with the main difference being the logistic function that is fitted to the training instances, instead of a linear function. To be able to fit a logistic function to the training instances, a simple least-squares method as used in linear regression is not applicable. To make this possible, the

probabilities p on the y -axis have to be converted to the respective *logit* (p) or log odds before, according to Eq. (6).

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \quad (6)$$

Equation 6 Calculation of *logit* (p) for curve fitting in a logistic regression

The y -axis now ranges from negative infinity to positive infinity, instead of zero to one as before. This allows one to draw a random linear line through this coordinate system. The training instances of both classes can be projected onto that line, since their x coordinate is still known (continuous feature x), and their respective *logit* (p) can be seen. Now that the *logit* (p) of each training instance is known, they can be converted back to probability values p . This transformation was necessary to calculate p of the training instances when using this particular logistic function, which were before just zero or one. Now, the $\log(p)$ values can be summed up yielding the log (likelihood), a quantitative measure of how well the logistic function fits the data points produced by the training instances. The algorithm of logistic regressions varies the linear line on the *logit* (p) plot in a way that log (likelihood) is optimized iteratively, and thus finds the best fitting line similar as to linear regression. There are several variants of logistic regressions known today. In the context of target prediction, multiple logistic regression is worth mentioning. It is used by the SwissTargetPrediction web server (Sect. 3.3). A multiple logistic regression can utilize several attributes as, for example, the SwissTargetPrediction uses two-dimensional similarity and three-dimensional similarity to calculate an overall probability.

Another popular and fairly simple category of MLAs for classification are the k -nearest neighbor algorithms, schematically described in Fig. 6c (in the literature often found as k -NN). The k -NN are group of algorithms that classify complex data points according to their immediate vicinity. To do so, the distance of each instance in a training dataset to every other instance in the same training dataset has to be computed. Additionally, every instance is assigned to one of the two possible classes “active” and “inactive.” For a new instance to be classified (e.g., test set), all distances to every training instance are calculated and the k -nearest neighbors selected; k is a natural number \mathbb{N} and stands for the number of nearest neighbors to be considered by the algorithm. So, in a 1-NN model, the algorithm picks the nearest neighbor to the instance to be classified and assigns it to that same class. In a 3-NN model, the algorithm first picks the three nearest neighbors to the instance to be classified, and assigns the class that is the most abundant in those instances. For example, if one instance belongs to the class “active” and two instances belong to the class “inactive,” the new instance will be classified as “inactive.” The value of k should be chosen depending on which performs best. General considerations are that a small k value like in 1-NN tends to have a poor signal-to-noise ratio, and predictions can be influenced heavily by, e.g., outliers, since only a few samples are considered. Large k values give more accurate predictions, but over-prioritize classes

containing many training instances. Comparatively smaller classes will constantly be overruled in favor of larger classes. For example, a 1-NN model is used within HitPick (Sect. 3.4).

In contrast to naïve Bayesian models, SVMs classify a population by discriminating projected data points in a vector space. Each data point is represented as a vector. The SVM then builds a hyperplane, aiming to achieve an optimal separation between vectors of different classes, e.g., “active” and “inactive.” Thereby, SVMs endeavor to maximize the distance between the hyperplane and the frontline of closest vectors of both classes (best possible separation). Vectors that are farther away from the hyperplane are neglected during this process. However, hyperplanes are actually linear, but most classification problems in the real world are not linearly separable. To account for this limitation, SVMs provide various “kernels” that allow the use of nonlinear hyperplanes (see Fig. 6d), making SVMs versatile and practically meaningful. The SVMs are, next to classifiers, also available as regression models, allowing for a continuous prediction of the dependent variable. For example, in admetSAR (Sect. 3.5), five SVM regression models are implemented. The VM classification models are used, e.g., in admetSAR (Sect. 3.5) and Pred-hERG (Sect. 3.6).

Next to the four MLAs presented, a myriad of other supervised learning methods are available and have been discussed elsewhere, like decision trees, random forests, AdaBoost, artificial neural networks, and deep learning. As indicated earlier, the fields of artificial intelligence and machine learning are currently developing at enormous speeds, as the general understanding of artificial intelligence and computational power keeps on growing. Leading tech companies are investing large amounts of resources into this technology, and new applications are being generated at a high rate. The benefits that come along with these efforts and are being made currently within and outside the drug discovery world, will certainly aid VS technology on the long run. Calculation times will be shortened and predictions will become more accurate. It remains to be seen, however, if artificial intelligence will revolutionize various fields of drug discovery, including target prediction. Applications of MLA classifiers for hERG prediction can be found, for example, in [76–89].

2.1.4 Ligand-Based Pharmacophore Modeling

A pharmacophore is a three-dimensional arrangement of physicochemical features that represent the interactions between a small molecule ligand and its protein target. It consists of hydrogen-bond donors, hydrogen-bond acceptors, aromatic features, hydrophobic features, positive ionizable features, negative ionizable features, and exclusion volumes. A pharmacophore model can be used to screen virtually large three-dimensional databases for other structures that fulfill the interaction pattern and consequently have a high probability of activity. Pharmacophore models either can be derived from crystallized ligand-protein complexes (structure-based approach; see Sect. 2.1.5.5) or from a training set of compounds with measured activities.

The structure-based approach produces models that closely reflect the biologically active conformation of a molecule within the binding pocket, featuring known key interactions. However, not all protein targets have been co-crystallized with a ligand molecule and many protein targets can be inhibited in more than one way. Ligand-based pharmacophore modeling can be used either if no crystallographic data is available or to complement structure-based models with models for alternative binding modes. To create ligand-based pharmacophore models, a small number of active and structurally similar compounds are aligned in three-dimensional to derive common physicochemical features. The shared feature pharmacophore model is then optimized by screening a training set database of known active and inactive molecules. Feature settings and placement are then optimized for maximal performance. Even though there is no steric information available from the protein, it is possible to include steric aspects by introducing exclusion volumes or a shape feature, based on the training set. The generation of ligand-based pharmacophore models is shown schematically in Fig. 7.

To cover all possible binding modes without losing model restrictivity, it is often feasible to perform a combined VS of multiple different models, each representing different classes of ligands (see Fig. 8).

Sidelining the difficulties of conducting structure-based modeling with homology models, ligand-based pharmacophore modeling has been used successfully in several instances for hERG activity prediction. In 2002, Cavalli et al. created a ligand-based pharmacophore based on QT-prolonging activity and combined it with a 3D-QSAR model for hERG channel blocking activity [90]. Durdagi et al. derived ligand-based pharmacophore models in PHASE based on 31 hERG channel blockers [91]. Aronov et al. proposed a pharmacophore model for charged [92] and uncharged hERG blockers in 2006 [93]. Pharmacophore modeling was also combined with descriptor-based models predicting hERG channel blockage [94]. Yamakawa et al. used pharmacophore modeling to describe the different structural requirements for hERG channel blockers and facilitators, compounds that block the channel but also enhance channel activation after the application of a depolarizing voltage step [95]. Kratz et al. published a set of seven ligand-based pharmacophore models for hERG channel blockers, based on 15 hERG blockers from the literature, validated both theoretically with literature data and experimentally by a prospective screening of large compound databases and subsequent biological testing. The models in this approach achieved hit rates of up to 66%, illustrating that pharmacophore modeling can be a powerful predictive tool for hERG-based cardiotoxicity [96]. In 2017, Durdagi et al. developed pharmacophore models for hERG-1 activation based on 18 hERG-1 activators from the literature [97].

2.1.5 Generation of Protein Structures for Molecular Modeling

As explained above, structure-based models are derived from experimentally solved protein-ligand complexes. In drug discovery, the three-dimensional representation of

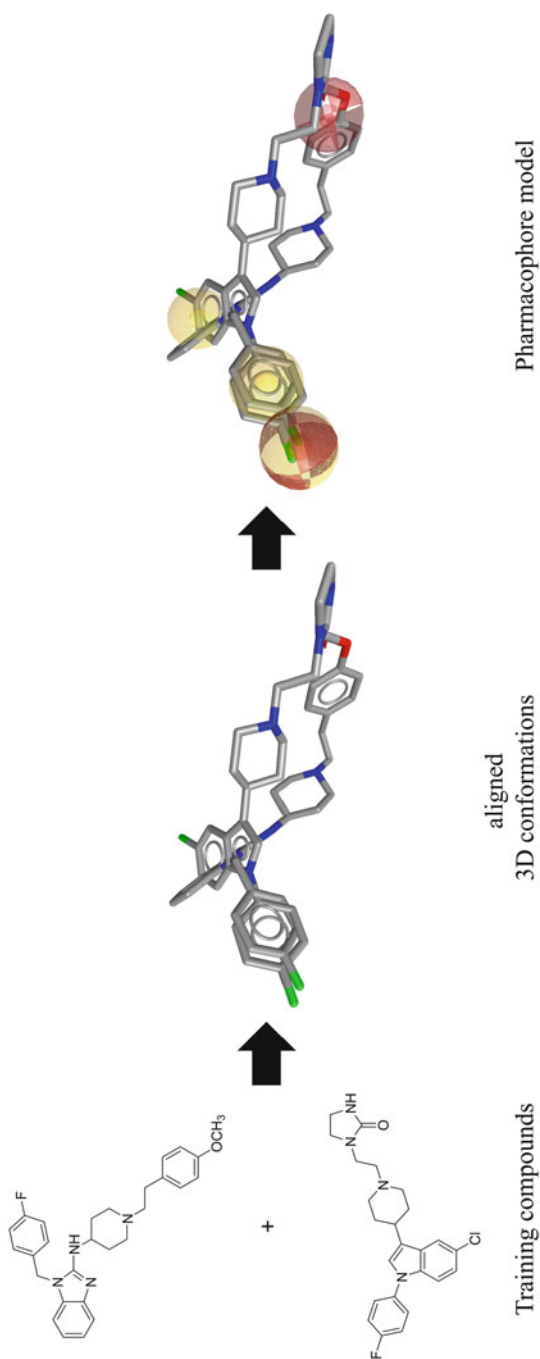


Fig. 7 Ligand-based pharmacophore model, generated from the hERG blockers astemizole (**9**) (top) and sertindole (**8**) (bottom). Three-dimensional conformations for the training compounds are computed, and the low-energy conformations aligned. From there, matching functionalities are summarized into a “shared feature” pharmacophore model. Red spheres, H-bond acceptor; yellow sphere, hydrophobic feature

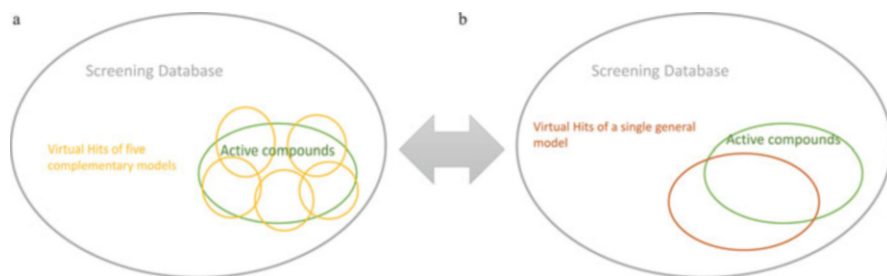


Fig. 8 Ligand-based pharmacophore modeling often requires the generation of several pharmacophore models that are meant to be used in cooperative mode. Complementary models can cover a larger chemical space of the true active compounds (left side) than just one comprehensive model (right side)

such complexes is most frequently obtained by X-ray diffraction measurements. Next to protein crystallography, NMR methodology is used to determine three-dimensional structures of proteins. More recently, yet another promising technique has been included in drug discovery, namely, cryo-EM. In the case of hERG, there is only one structure available, which was solved using cryo-EM by Wang and MacKinnon in 2017 [10]. In the following paragraphs, an overview of the three abovementioned experimental techniques will be provided, explaining their respective constraints and strengths. Also, a theoretical technique to obtain three-dimensional structures of proteins will be described, which has played a major role in hERG research over the last decade: homology modeling.

2.1.5.1 X-Ray Crystallography of Proteins

X-ray crystallography is not just the oldest of the three experimental techniques mentioned above, it is also the most important technique in drug discovery. The roots of X-ray crystallography date back to experiments of Max von Laue in 1912, in order to prove the wave character of X-rays, and his findings are hence the very foundation of X-ray crystallography: Crystalline matter diffracts X-rays. The diffracted X-rays can be measured, or to be more precise, the angles and intensities of the diffracted beam can be used, to deduce the actual electron density of the sampled crystal. The electron density is located around the atoms of the crystal, and thus the electron density map can be seen as an empty shell around the protein's molecular skeleton. However, a crystal under investigation is never a single protein, but a large agglomerate of proteins. This agglomerate in the crystal is made up by a highly ordered crystal lattice that can be described by the primitive cell. Therefore, electron density maps resulting from X-ray crystallography depict the primitive cell as a mean over all the measured primitive cells. Ultimately, this can be used to model the protein sequence inside. Protein structures solved by X-ray crystallography are

typically not that well resolved to permit a direct assignment of atoms. It is much more feasible that the protein sequence is modeled into an electron density map.

For example, since hydrogen atoms only have one electron, hydrogen assignments are typically set empirically. In a protein, the most common hydrogen species is connected to a carbon, for which typical angles are known. Hydrogen atoms that are covalently bound to heteroatoms are more flexible, but here a contextualization into an overall H-bond network can help. Nevertheless, the better the resolution of a solved crystal structure is, the easier the assignment. Usually, resolutions below 2.5 Å are considered as well resolved (e.g., suitable for molecular docking). In contrast, electron density maps that were not well resolved result in structures that are potentially erroneous, at least in some regions, to a certain degree. Such structures are eventually unsuited for techniques like molecular docking. Cole et al. reviewed extensively the quality attributes of structures produced by X-ray crystallography in regard to molecular docking [98]. However, the narrowest bottleneck in X-ray crystallography is the crystallization of the protein. Biomolecules in general do not crystallize very well, and thus, many decades of development work have been necessary to overcome this hurdle.

The advantage of X-ray crystallography is definitely its ability to reach very good resolutions, and to be able to depict large proteins such as receptors. The biggest drawbacks still lie in the crystallization procedures themselves, since proteins, like those bound in the membrane, crystallize very poorly. This on the one hand is due to their dynamic behavior and on the other hand as a result of their large hydrophobic surfaces turned to the outside (interaction sites for membrane). Moreover, the crystallized state is not necessarily a realistic projection of the *in vivo* conformation. Even further, proteins are *per se* dynamic, meaning that a single conformation is often not enough to properly describe a “living” protein accurately. However, X-ray crystallography has contributed greatly to many disciplines over the last decades, and its insights have fueled numerous successful projects in drug discovery. From the 137329 protein structures deposited in the RCSB-PDB, 123978 (90.27%) were solved by X-ray crystallography [99]. Over the last 10 years, the number of three-dimensional structures deposited in the RCSB-PDB has grown constantly each year, indicating that X-ray crystallography is still of the utmost importance [11–14].

2.1.5.2 Biomolecular NMR

Biomolecular NMR is yet another technique to solve three-dimensional structures of proteins. While the technique of nuclear magnetic resonance (NMR) has for decades been the gold standard method for structure elucidation of small organic compounds, the latest advances in the field of NMR have focused on large biomolecules. Biomolecular NMR uses the exact same principles as classical NMR, with the difference that proteins are considerably larger than small organic compounds. Moreover, proteins are built up by only 20 units of structurally quite similar amino acids, meaning that, e.g., a regular ^1H or ^{13}C spectrum becomes very crowded by overlapping peaks. To tackle this problem, multidimensional NMR experiments

have been developed that allow for an assignment of the respective chemical shifts to a distinct atom—the hallmark for any NMR data evaluation. These experiments were pioneered mainly by Richard R. Ernst and Kurt Wüthrich, who were awarded the Nobel Prize in Chemistry in 1991 and 2002. Biomolecular NMR is limited inherently to five types of nuclei, which possess a spin and are present in a biomolecule: ^1H , ^{13}C , ^{15}N , ^{19}F , and ^{31}P , whereas only the first three are actually of relevance in proteins. However, such multidimensional experiments can aid the assignment of NMR peaks. Moreover, Nuclear Overhauser Effect Spectroscopy (NOESY) can give insights into non-covalent atom interactions, like those formed through protein folding. The sum of such information allows for the construction of several three-dimensional structures of the protein.

The major advantage of biomolecular NMR is that data are acquired from solution, which, in general, represents an environment closer to reality for proteins than the crystallized state. Moreover, dynamic movements of the protein can be observed. The major disadvantage without doubt is its limitation to relatively small proteins. These limitations result from several factors, like the obvious increasing number of peaks in the spectrum, or the shorter relaxation times. Moreover, biomolecular NMR structures are typically an ensemble of possible structures. From the 137,329 protein structures deposited in the publicly accessible RCSB-PDB, 10,962 (7.98%) were solved by biomolecular NMR [99]. However, the number of such structures that were deposited each year in the RCSB-PDB has diminished slightly over the last decade [11–14].

2.1.5.3 Cryo-Electron Microscopy

In 2017, the Nobel Prize in Chemistry was awarded to Jacques Dubochet, Joachim Frank, and Richard Henderson to honor their joint efforts in cryo-electron microscopy (cryo-EM). Even though the foundations of this electron microscopy technology date back to the 1930s (Max Knoll and Ernst Ruska), it was not possible to image biological samples for a long time and very difficult to reach atomic resolution. Recent efforts have finally led to a boost in high-resolution cryo-EM images of proteins, mainly due to the improving instrumental capacities and image processing software. Moreover, the software to run the respective instruments has become more user-friendly and the technique as a whole more established, allowing its application now also to non-EM specialists. The process of a cryo-EM model generation roughly can be divided into five steps: first, the protein of interest needs to be expressed in a sufficient amount and sufficient purity. Second, the sample needs to be prepared, so that the protein is kept in a natural or near-natural conformation while simultaneously optimizing the image contrast. The first requirement can be achieved such as by optimizing the buffer system or by stabilizing membrane proteins using amphipols or detergents [100]. The second prerequisite can be achieved by, e.g., removing sugars or glycerol. Third, the sample needs to be fixed onto a grid to be introduced into the instrument. Currently, this is usually done by placing the sample solution on the sample grid and subjecting the loaded grid to rapid cooling in liquid

ethane. This process is also referred to as “vitrification” and turned out to be of pivotal importance to ensuring the most recent successes in cryo-EM. The flash cooling leads to the formation of amorphous ice around the sample, which, in contrast to crystalline ice, does not interact with electrons. This thus behaves like transparent glass holding the protein sample in the conformation it had in the buffer solution. Moreover, the low temperature protects the sample from the relatively strong electron beam emitted by the electron microscope. Fourth is the actual image acquisition, in which the flash-cooled grid is placed into the transmission electron microscope and several thousand images are taken. The image acquisition in cryo-EM is quite elaborate and can take up to a few days. This shortcoming mainly results from the fact that the contrast between the sample and its background (such as buffer) is still relatively low, although having been significantly increased over the last years. Therefore, it remains necessary to acquire many single images and to average these computationally in order to reach meaningful contrasts. Fifth, the image post-processing stage plays a major role as well. To point out its central importance, it should be noted that enhancements in this step especially have boosted cryo-EM technology. Algorithms employed for this particular task are able to detect the various protein particles in the sample, classify their relative orientation and the projected angles (such as top view, side view, side view 30°), and to reconstruct a three-dimensional model out of those various two-dimensional images. The resulting three-dimensional model further can be refined manually, to obtain the best possible resolution.

The obvious advantage of cryo-EM over X-ray crystallography is its applicability in resolving large and very large complexes or assemblies while being independent of the crystallization process. For example, efforts to crystallize human hERG have failed so far. Moreover, vitrification allows for the detection eventually of various possible protein conformations, while X-ray crystallography typically covers just one conformation. The conformations in cryo-EM are near native, since they are fixed while being in buffer solution. In X-ray crystallography, however, the one detected conformation is, by design, typical for the protein in solid state, which does not necessarily reflect its native conformation. Therefore, cryo-EM also enables the imaging of dynamic, flexible proteins like, e.g., G-protein-coupled receptors [101]. The clear disadvantages of cryo-EM are however the long acquisition times and the currently still lacking automation of the process. Sample preparation still needs to be performed manually. For example, Renaud et al. state that in “high-throughput crystallography,” data of 500 crystals soaked with different compounds can be acquired in just 24 h and post-processed approximately in a few days. In cryo-EM, measuring the same amount of samples would take approximately 1.5 years [102]. Another disadvantage is the relatively low resolution still produced by cryo-EM compared to other techniques, as demanded by, e.g., molecular docking [103]. For example, the hERG structure reported by Wang and MacKinnon is resolved at 3.7 Å, which is not far away from the threshold set for docking experiments (2.5 Å) [98]. However, 3.7 Å ranges in the magnitude of two to three C-C single bonds, which makes molecular modeling studies with this structure blurry, but not impossible. This situation very likely will improve over the next

few years, as technology keeps on progressing. Cryo-EM is, after all, the fastest growing technique in this field, with the greatest increases in entries per year in the RCSB-PDB since 2010. As for January of 2019, of the 137,329 proteins listed in the RCSB-PDB, 2016 (1.47%) were solved by cryo-EM [99]. This number doubled in just 2 years. In addition to the RCSB-PDB, cryo-EM data are also deposited in the EMD.

2.1.5.4 Homology Modeling

Whenever a three-dimensional structure of the desired protein is not available, homology modeling represents another option, next to pure structure-based and pure ligand-based modeling. In short, it is possible to construct a three-dimensional model of the desired protein using a closely related protein as template, for which the structure has been solved. Closely related means that the respective amino acid sequences should be considered as homologous, with not less than 30% identical residues. Homology models with more than 50% sequence identity usually produce reliable models, even though particularly flexible parts of the protein (e.g., loops) are prone to errors [104–106] and the active center is often the very point of distinction between closely related proteins. Thanks to the Basic Local Alignment Search Tool (BLAST), the complete PDB can be sampled in relatively short time span [107]. Homology models can ultimately be seen as models of a protein for which the structure is not yet available. Homology models can also be used, e.g., for molecular docking, to derive structure-based pharmacophore models, for molecular dynamics simulations, and other purposes. Thus, homology modeling is a technique to model a protein's three-dimensional structure, which then allows for structure-based methods. Regarding hERG, several homology models have been proposed prior to the first cryo-EM structure in 2017 [23, 108–111]. This reflects the great interest of researchers for hERG as an antitarget.

2.1.5.5 Structure-Based Pharmacophore Modeling

Structure-based pharmacophore models rely on the availability of crystallographic data on the protein-ligand complex. While it is easy to generate a model based on protein-ligand complexes, automatically calculated models can almost always be improved by optimizing them with a training set to maximize active enrichment in their hit lists.

As discussed above, a three-dimensional structure of hERG has been published quite recently, so there are not yet many publications based on this new information. Nevertheless, examples for structure-based modeling using this hERG structure are, e.g., [112, 113], and Munawar et al. who derived structure-based pharmacophore models [114]. The generation of structure-based pharmacophore models is described schematically in Fig. 9.

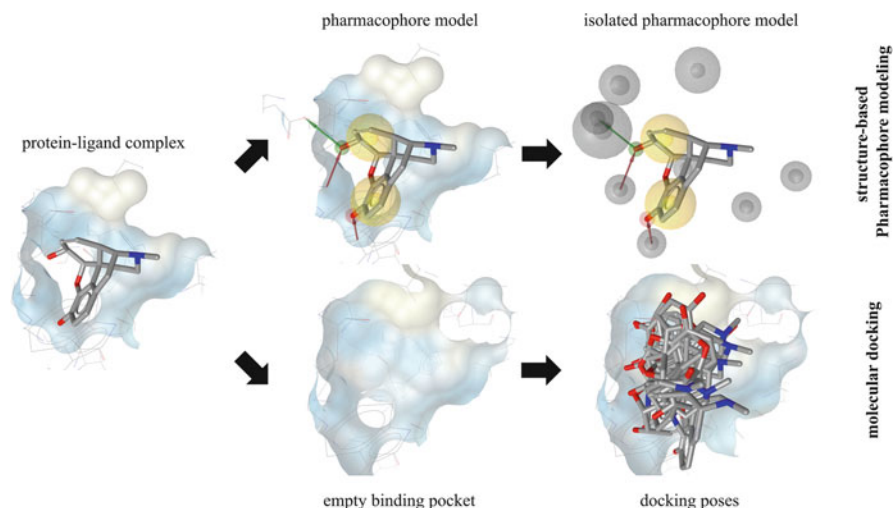


Fig. 9 Examples of structure-based modeling with a protein-ligand complex from hERG. In structure-based pharmacophore modeling (top), a pharmacophore model is generated from a protein-ligand complex, and extracted. Amino acid residues are finally converted to the so-called exclusion volumes (gray spheres), forbidden zones that would lead to spatial clashes. In molecular docking (bottom), the native ligand is removed from the binding site, before the screening ligand (s) are placed into it. The conformational space of the ligand(s) is explored inside the protein's binding pocket in an iterative manner, and a scoring function is applied. The result is a set of ranked docking poses for each screened ligand. Images were created using LigandScout 4.2 (Inte:Ligand, Vienna, Austria) and the PDB entry "5VA1" [10]

2.1.6 Molecular Docking

Molecular docking is perhaps the most popular and possibly even the most well-known technique in VS and iVS currently. Since molecular docking is described extensively elsewhere in the literature [115–121], its theoretical foundation will only be described rather concisely below. In general, molecular docking is a completely structure-based approach, meaning that an experimentally solved or theoretically generated three-dimensional representation of the desired protein is inalienable. A docking run can always be described in two phases: the actual “docking” and the “scoring.” During the docking, an algorithm samples a vast amount of ligand conformers inside the binding site, while during the scoring, the hypothetical energies of the hypothetical protein-ligand complexes are computed. However, the processes of docking and scoring occur in an alternating manner, so that each complex generated is scored immediately after being docked. The algorithm repeats this process many times, while trying to constantly minimize the calculated energy of the hypothetical protein-ligand complexes at each iteration. Within docking, the ligands in such complexes are referred to as “docking poses.” Finally, the docking software can output either the best-scored (lowest energy) pose or a list of poses,

ranked according to their respective score. Molecular docking is shown schematically in Fig. 8.

These principles are true for all docking techniques, regardless of which subcategory of docking is being used (with the exception of protein-protein docking). Typically, docking techniques are divided into how protein flexibility is handled: If the protein and the ligand are kept flexible, the correct term is “flexible docking.” Unfortunately, flexible docking requires considerable computational efforts but is still used in drug discovery. Usually, VS campaigns aim to screen hundreds or even thousands of compounds or targets, making time-consuming applications unfeasible. To tackle this shortcoming, VS is usually conducted using “rigid body docking.” In rigid body docking, the protein is assumed to be rigid, and only ligand conformations are sampled. This simplification greatly reduces calculation times and enables even high-throughput VS. The assumption that a protein behaves like a rigid body through ligand binding is of course incorrect. Flexible docking is much closer to reality, but still rigid body docking can produce good results. Moreover, a middle ground of these two techniques exists, namely, “induced-fit docking.” In induced-fit docking, parts of the protein are considered as flexible, e.g., certain amino acid residues or a loop. Induced-fit docking represents a compromise between the higher accuracy that may be attained through flexible docking and the increased speed of rigid body docking.

The scoring during a docking run is conducted by a so-called scoring function—a predefined ruleset to evaluate the meaningfulness of the generated poses. It has been a constant endeavor in the development of the docking technology to implement a scoring function that accurately computes the ligand’s affinity to the proteins to which it is docked. However, such a scoring function has still not been found. They are either computationally too expensive to use in a docking run or too simplified and thus lack accuracy. Scoring functions that are in place today are designed instead to produce relative scores for each compound. Based on their working principle, they can be subdivided into knowledge-based, empirical and force field-based. Knowledge-based scoring functions are derived from large databases containing experimentally solved protein-ligand complexes. From these, the “knowledge” of typical protein-ligand interactions in terms of nature, distances, and angles can be mined and translated into a scoring function. Empirical scoring functions are derived from databases containing a multitude of experimentally assessed binding affinities together with the respective structural features of the protein-ligand complex. Such relationships can be used as training instances, for the construction of a statistical model, which can extrapolate this knowledge to new compounds. These models can include linear regressions but also more sophisticated MLAs like, e.g., SVMs or random forests. Scoring functions have been reviewed in great detail in [116, 122, 123].

A docking workflow practically consists of three main steps. First, the inputs have to be prepared—protein and the ligand(s). An appropriate three-dimensional representation of the protein has to be found, the binding site defined, eventual ligands removed, as well as optional modeling of the protein conducted (e.g., minimization of loops, flipping of side chains). The ligands have to be selected, three-dimensional

starting conformations computed, as well as protonation states and tautomeric states generated. The second step comprises the actual docking procedure as described above (consisting of docking and scoring). The third and final step is the post-processing of the results. This can include, e.g., rescoring of the poses using a different scoring function, applying pharmacophore models to screen for known binding modes, or using shape filters to introduce specific steric requirements. However, post-processing is highly customizable and depends on the inputs used, the results obtained, and the goal of the docking. In terms of computation time, molecular docking is a relatively costly technique used in VS and definitely the most expensive general method described in this chapter. By rule of thumb, the computation times rise in the following order: 2D ligand-based similarity searches = machine learning-based similarity searches < approximate shape-based searches = pharmacophore searches < Gaussian-based shape searches << molecular docking.

For a long time, structure-based approaches for the modeling of hERG were not possible. Therefore, the only structure-based approach featured in this study is VirtualToxLab™ (see Sect. 3.8). VirtualToxLab™ uses a modified induced-fit docking algorithm, based on the homology model proposed by Farid et al. [109]. It is very likely that this situation will change very soon, based on the recent advances made by Wang and MacKinnon [10].

2.2 Evaluating Target Prediction Models

To assess the overall predictive power of each tool under investigation, our group computed receiver operating characteristic (ROC) curves and calculated the following quantitative metrics: true positive hits (TP), false-positive hits (FP), true negative hits (TN), false-negative hits (FN), precision (*Pr*), sensitivity (*Se*), specificity (*Sp*), F-measure (F_1), accuracy (*Acc*), and the area under the ROC curve (ROC-AUC).

In VS, a TP describes a virtual hit that was validated experimentally, while an FP is a virtual hit that does not show the predicted activity in the experiment. A hit list therefore typically consists of TPs and FPs. True negatives are accordingly compounds that are predicted to be inactive by the model, which can be proved experimentally. False-negative hits are compounds that are predicted as inactive by the model, but show activity in vitro.

The metrics TPs, FPs, TNs, and FNs are simple numbers, describing the amount of compounds belonging to each class. However, these numbers are mainly dependent of the size of the screened database, and thus do not allow for an objective judgment at first sight. One might understand that not the absolute number of these four metrics, but rather the ratio between them is of interest in describing a model's performance. As TP, FP, TN, and FN reflect numbers from zero to one, all of the following metrics that are calculated from the latter also range from zero to one. One always indicates the highest reachable score, while zero always indicates the lowest reachable score.

One of these measures is Se , or often referred to as “recall” or “true positive rate” (TPR). The Se describes how many true actives a model can retrieve from a screening database [Eq. (7)]. Sp , on the other hand, often referred to as true negative rate (TNR), describes how many inactives can be retrieved from a screening database [see Eq. (8)] [124].

$$Se = TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (7)$$

Equation 7 Calculation of Se . P: in vitro active molecules

$$Sp = TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \quad (8)$$

Equation 8 Calculation of Sp . N: in vitro inactive molecules

In general, a model with only high Se or Sp does not necessarily produce good results, since both of these metrics only consider one side of the activity spectrum (retrieval of active or inactive hits from actives or inactives, respectively). Therefore, Pr is yet another popular metric. The Pr , or also known as the “positive predictive value” (PPV) or the “yield of actives” (Ya), describes the amount of TPs in a hit list. The Pr is a measure of how likely it is that a hit in a hit list is really active [see Eq. (9)] [124, 125].

$$Pr = PPV = Ya = \frac{TP}{n_{hits}} = \frac{TP}{TP + FP} \quad (9)$$

Equation 9 Calculation of Pr . n_{hits} : number of hits

Also, Pr only takes into account the actual hit list, and not so much the retrieval of actives or inactive from the screening database. For example, a model could have a high Pr but a low Se , meaning that most of the compounds in the hit list are in fact active, with many actives being misclassified as inactive by the model. The values Se and Pr are very similar in terms of their calculation, even though their meaning is different. A model could be very sensitive by correctly retrieving all actives from the screening database but still lack in precision by producing also a high number of FPs. Both cases just described would actually make this model not very reliable, even though single quantitative metrics seem appropriate at first sight. Thus, it is desirable to have more than just one metric close to one, e.g., Se and Pr , or Se and Sp , etc.

To properly evaluate a model, more global, objective metrics are necessary: Acc and F_1 equally can be computed from the four initial metrics TP, FP, TN, and FN. The F_1 score is a metric incorporating both of those prior metrics, by combining Se and Pr into their harmonic mean. Therefore, one can expect from models with an F_1 score close to one to have both good Se and Pr [see Eq. (10)]. The Acc , on the other hand, describes the overall correctness of the model’s classification results, by

comparing all true predictions (TP and TN) with the overall entries in the screening database ($n_{compounds}$), like shown in Eq. (11) [125, 126].

$$F_1 = 2 * \frac{Se * Pr}{Se + Pr} \quad (10)$$

Equation 10 Calculation of F_1

$$Acc = \frac{TP + TN}{n_{compounds}} = \frac{TP + TN}{P + N} \quad (11)$$

Equation 11 Calculation of Acc

The ROC curve is a very popular metric to describe the predictive power of any binary classifier. The biggest advantage of ROC curves over other metrics is the 2D plot. An ROC curve can be visually easy to interpret and thus gives more clues about the actual performance behavior of a model. Other typical metrics only return single numbers that often do not meet the complexity levels of the actual performance behavior of a model. ROC curves are plots, where the Se is plotted as a function of the false-positive rate (FPR), which corresponds to $1 - Sp$. As explained above, Se is a direct measure on how many actives can be retrieved from the screening database. The FPR , on the other hand, describes how often FPs occur. The FPR therefore behaves complementary to the specificity, which describes how often inactives are correctly predicted as TNs. To simplify: the hit list is sampled hit by hit, starting from the most probable (highest scored) to the most improbable (lowest scored) virtual hit. If a hit is classified correctly (TP), the ROC curve directs upward ($TPR > FPR$). Misclassified hits (FP) make the ROC curve turn left ($TPR < FPR$). If $TPR = FPR$ over the complete dataset, so then the model would not seem to show any preference for true actives or true inactives. The model classifies on a random basis, which is no better than if one is rolling a dice. Models with good predictive power produce significantly higher TPRs than FPRs ($TPR \gg FPR$) and run above the “random line.” A bad model can even show ROC curves that run below that “random line” ($TPR \ll FPR$).

The ROC curve can also be integrated. The properties explained above have indicated that the ROC-AUC is proportional to the predictive power of the model. Since the ROC curve is plotted on a 2D coordinate system with x - and y -axis ranging from zero to one, the ROC-AUC also ranges from zero to one. One indicates a perfect model, with a TPR of one (the hit list contains only hits that are active) and a FPR of 0 (the hit list contains no inactive hits). In contrast, an ROC-AUC of 0 indicates a TPR of zero and an FPR of one, indicating that every hit in the hit list is misclassified. An ROC-AUC of 0.5 indicates a random model where there is the same amount of TPs than FPs in the hit list [124, 127]. The metrics describing the in silico tools used during this study are summarized in Sects. 4.4.1 to 4.4.3.

3 Publicly Available Target Prediction Tools Suitable for hERG

3.1 Similarity Ensemble Approach

The Similarity Ensemble Approach (SEA, <http://sea16.docking.org/>) is a publicly available web server developed by Keiser et al. at the University of California, San Francisco (UCSF) [128, 129]. SEA operates according to the above-described principle of 2D ligand-based similarity, by comparing a screening compound to reference compounds, which are each associated with a protein target. To perform such a task, a large annotated database is required, containing several compounds in a computer-readable format and its respective confirmed bioactivities. The Similarity Ensemble Approach uses the ChEMBL database, which is maintained and curated by the European Bioinformatics Institute (EBI), of the European Molecular Biology Laboratory (EMBL) [31–33]. The ChEMBL database contains a large collection of bioactive chemical entities with drug-like properties that are manually retrieved from scientific literature, making ChEMBL one of the most popular chemical databases for computer-aided drug design. For SEA, the ChEMBL data was processed and sorted according to protein targets, while chemical structures were translated into ECFP4 fingerprints [56]. For target prediction, a screening compound entered to the SEA server is first converted to an ECFP4 fingerprint, and subsequently compared to any compound from each subset stored in SEA. For each comparison, a T_c is computed and kept if the absolute value is greater than 0.57. In a third step, all the T_c scores kept are summed up for each protein target, yielding the so-called raw similarity score (RSS). Raw similarity scores are thus a measure of how many confirmed binders to protein X are similar to compound a subjected to the target prediction. Of course, this measure, and therefore the RSS as well, is inherently biased by the original dataset size of protein X . If protein X is represented by a 500-compound-large dataset, while the dataset of protein Y contains only 40 compounds, the probability of finding similarities between screening compound a and dataset X is higher than with a and Y . This correlation is valid independently of target class or chemical substructure. To tackle this internal bias, RSS values were subjected to a z-transformation, yielding a z-score Z_s . Moreover, RSS values calculated from n comparisons of random sets were computed and z-transformed, resulting in a z-score distribution of random = $\{Z_{R_1} \dots Z_{R_n}\}$, representing the distribution for similarity comparisons between random chemical structures. Thus, it is legitimate to calculate whether the resulting Z_s of screening compound a is significantly different from random. Moreover, expectation values, E , are computed, which are interpreted as follows: E indicates the number of hits one would observe by pure chance at the same Z_s or higher. More practically, it can be summarized that in SEA, a low E is associated with high similarity to the screening compound and known binders of the predicted target. Accordingly, SEA issues a table of predicted protein targets, starting from the lowest E (most probable) to the highest E (least probable). Already known targets of the respective screening compound are also

highlighted by SEA. This can be detected easily by the algorithm if a T_c of one is observed (indicated 100% structural similarity), which can only be achieved if two identical compounds are compared to each other.

The core concept of SEA was inspired by the BLAST algorithms [107], a group of algorithms for sequence alignment popular in bioinformatics, ever since it was a core desire in bioinformatics to be able to compare how similar the sequences of proteins or nucleic acids are. As the sheer amount of available sequences has grown, there has become an increased need for efficient algorithms that are fast enough to sample through these growing databases. First, the sequence is cut into smaller pieces (e.g., three amino acids, called “words”) that are matched with the database entries. Once a database entry is found that contains this word, this match is expanded continuously. This principle is called “local alignment” and enabled BLAST algorithms to sample through large databases a lot faster than its successors performing the so-called global alignments. Now analogously one could argue that SEA describes a protein target by the chemistry of its ligands, rather than a sequence of amino acids. The single-reference ligands of a protein target would then correspond to the “words” in a BLAST search. The input of a SEA search is already in a word-like format, being a potential ligand itself. Comparisons between the input compound and the reference ligands are computed, and only those with a similarity score higher than a predefined cut-off are considered and accumulated.

Accordingly, protein targets with many similar ligands will produce high similarity scores in SEA, while sequences with many similar “words” will produce high similarity scores in BLAST. Both approaches compare the resulting similarity scores with a background distribution, which would be obtained from random similarity comparisons. Since in both cases this random distribution follows an extreme value distribution, it is legitimate to calculate expectation values E like those shown above. Also the interpretation of E remains the same as for a BLAST search. The calculation of E -values for 2D similarity comparisons is outlined schematically in Fig. 10.

3.2 *SuperPred*

SuperPred (<http://prediction.charite.de/>) is another publicly accessible web server, developed by the structural bioinformatics group of the Charité—University Medicine Berlin, Germany [63, 130, 131]. SuperPred’s working principle is, similar to SEA, also derived from the BLAST, meaning that protein pharmacology is compared via the chemistry of their ligands. To do so, every protein target that can be predicted is represented by a set of reference ligands. Moreover, next to predicting druggable protein targets, SuperPred can also predict drug classes based on an ATC code. Since hERG is not a drug class, and is therefore not represented in the ATC code, the drug class prediction suite of SuperPred will only be briefly introduced in this contribution. For the purpose of target prediction, SuperPred was equipped with the protein-ligand interaction data derived from ChEMBL [31–33], BindingDB [132], and SuperTarget [133] to build the reference ligand datasets for every protein

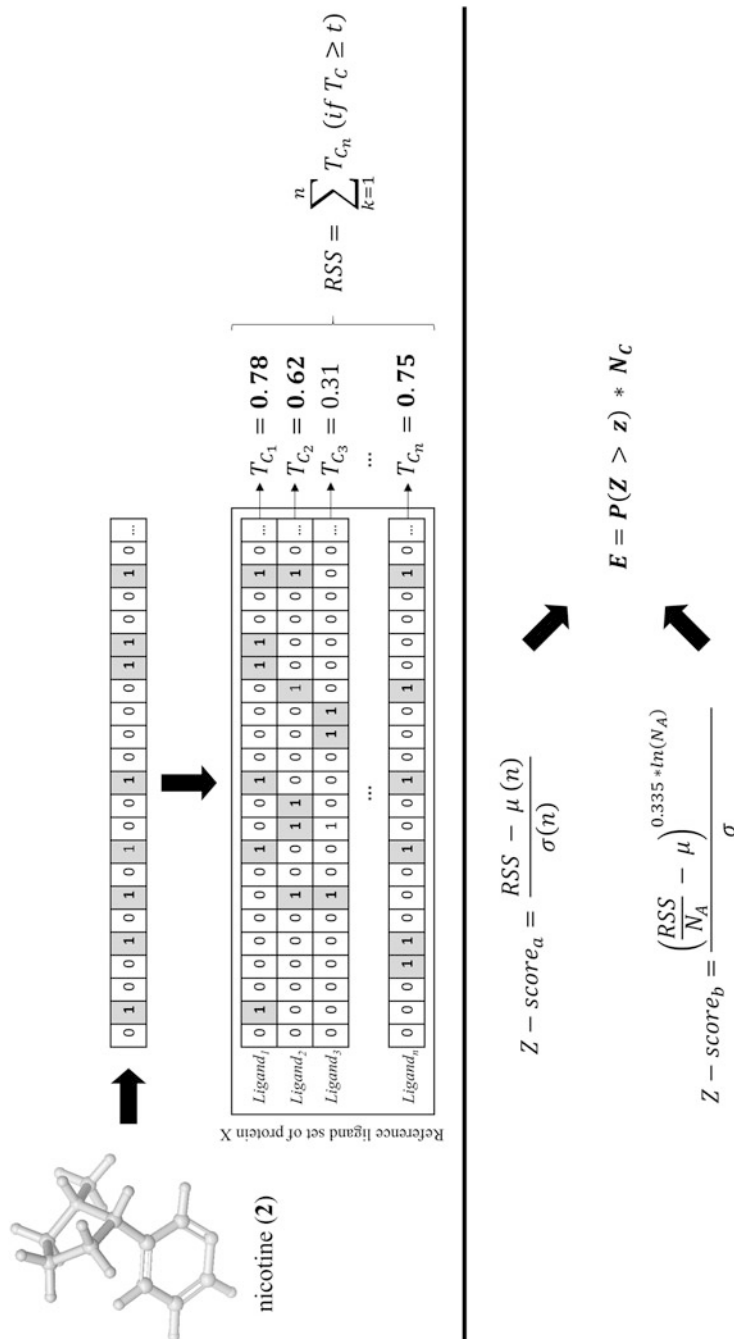


Fig. 10 Calculation of an E -value in the process of 2D similarity comparisons, using the hERG blocker **2** as example. Nicotine (**2**) is converted to a bit string using molecular fingerprints. $T_{c,s}$ are computed for all pairs of **2** and any hERG dataset entry. $T_{c,s}$ that are greater or equal to t are summed up to yield the RSS value. The RSS value is z -transformed, and compared to the z -score of a random distribution. Finally, E is calculated, by multiplying the probability that the obtained z -score can be reached only due to chance $P(Z > z)$ with the number of total set comparisons made N_c .

target. Similar to those mentioned above for SEA, SuperPred also uses ECFP4 fingerprints [56] for describing molecular structures. A compound submitted for target prediction to the SuperPred server is then first translated into a bit string according to the ECFP4 pattern. Second, the compound is compared to every reference compound in SuperPred and the respective T_{cs} are computed. T_{cs} above 0.45 are summed up to yield RSSs. Since RSSs, as shown by Keiser et al. [128, 129], correlate linearly with the size of the reference ligand set, RSSs need to be normalized to give meaningful results. In SuperPred, this is done by a division of the RSS by the number of reference ligands in the respective reference set. To the normalized RSS, a weighting factor λ is multiplied to account for different intra-reference set assemblies. More diverse reference sets are prone to yield lower RSSs, while less diverse reference sets can produce higher RSSs. If one reference set is composed for a major part of one series of ligands (which are structurally closely related), a screening compound that is similar to that series will yield many T_{cs} above 0.45 and consequently produce higher RSSs. Finally, this normalized, corrected RSS is compared to the background distribution of random set comparisons, similar to information outlined above for SEA and BLAST, yielding the z -score. Starting from the z -score, expectation values, E , can again be calculated for each predicted interaction. Low E -values correspond to high probability of the interaction not to be predicted by pure chance. The calculation of E -values for 2D similarity comparisons is outlined schematically in Fig. 9.

While the target prediction suite of SuperPred relies on the principle of 2D similarity, the ATC code classification is making use of various ligand-based approaches. Again, like in the target prediction suite, the 2D similarity is quantified by calculating T_{cs} from ECFP4 fingerprints. Furthermore, SuperPred compares the three-dimensional similarity by a Gaussian shape method and the 2D similarity of the respective compound's fragments.

3.3 *SwissTargetPrediction*

The SwissTargetPrediction server was developed by the Swiss Institute of Bioinformatics (SIB) and published by Gfeller et al. in 2014 [134, 135]. SwissTargetPrediction can be accessed from the SIB website (<http://www.swisstargetprediction.ch/>) and is provided free of charge. SwissTargetPrediction's idea is based on the observation that 2D ligand-based similarity often performs better within high structural similarity, while three-dimensional ligand-based similarity searches tend to perform better within low structural similarity. Thus, the SwissTargetPrediction paradigm was developed to take advantage of both strengths, by calculating two-dimensional and three-dimensional similarities between the screening compound and various reference ligand sets and combining the latter. The combination of two-dimensional and three-dimensional similarity scores is achieved over a logistic regression, a supervised MLA used for classification. The SwissTargetPrediction relies on a solid data foundation, representing all predictable protein targets by a representative set of

confirmed ligands. For the SwissTargetPrediction, similar observations were seen with SEA and SuperPred, and these data were derived from ChEMBL [31–33] and processed as described in Gfeller et al. [134]. Each ligand in every reference ligand set of a protein target is stored in two representations: first, as FP2 fingerprint, as implemented in OpenBabel (version 2.2.0) [136], accounting for the two-dimensional similarity. Second, as five-dimensional vector (x-axis, y-axis, z-axis, partial charges, $\log P$) for a maximum of 20 low-energy conformations per ligand, accounting for the three-dimensional similarity.

Whenever a screening molecule is submitted to the SwissTargetPrediction web server, two-dimensional and three-dimensional similarities are computed in parallel. Two-dimensional similarities are computed by first translating the screening molecule into a bit string, according to the FP2 fingerprint scheme, and subsequently calculating T_c s describing molecular similarities to all reference ligands. Similar to other approaches, a threshold of 0.25 for minimum similarity was chosen, for which values below that threshold are regarded as dissimilar. All the similarity comparisons yielding a $T_c > 0.25$ are stored. Three-dimensional similarities are computed according to the principle of USR, described above. First, 20 low-energy conformations of the screening compound are generated and translated to a five-dimensional vector (identical procedure as for the reference ligands). Next, the vectors of all of the 20 conformations are compared to all vectors of the reference ligand and the Manhattan distance calculated. As indicated in Sect. 2.1.1, the Manhattan distance can be converted into a similarity measure S_{3D} , ranging from 0 (maximum dissimilarity) to 1 (maximum similarity, identity) according to Eq. (1). For three-dimensional similarity comparisons, a cut-off of 0.65 was introduced, classifying ligand pairs with $S_{3D} < 0.65$ as not significantly similar. Finally, both T_c and S_{3D} values may be normalized between zero and one and used as input features for a multiple logistic regression model (as described in Sect. 2.1.3). The logistic regression, being an MLA, “learned” the optimal weighting of T_c and S_{3D} during a training phase to yield the most accurate predictions. This model may be applied subsequently for posing target predictions of novel ligands.

3.4 *HitPick*

HitPick is another fully ligand-based target prediction tool that can predict hERG interactions (<http://mips.helmholtz-muenchen.de/hitpick/cgi-bin/index.cgi?content=targetPrediction.html>). It was developed by the Institute of Bioinformatics and Systems Biology and German Center for Diabetes Research of the Helmholtz Zentrum in Munich. HitPick is available free of charge, and its functionality is discussed in detail in [137]. Recently, HitPick was updated to HitPickV2; however, the present study was conducted with the first version of HitPick [138]. HitPick combines two 2D ligand-based similarity methods, both using MLAs, namely, INN and a Laplacian-modified naïve Bayes algorithm (both outlined in Sect. 2.1.3). Typical for 2D ligand-based approaches, also HitPick bases its predictions on reference ligands that are

annotated with their respective, confirmed protein target. Here, the dataset was derived from the STITCH 3.1 database [139] and processed as described in the respective publication. The mined STITCH 3.1 database was ultimately used to build and train a 1-NN model and a Laplacian-modified naïve Bayes model using the FCFP as input [56].

A screening compound that is submitted to the HitPick server is then first translated into an FCFP-like fingerprint, like implemented in the RDKit (<http://rdkit.org>), an open-source cheminformatics toolbox, and stored. The nearest neighbor of the stored fingerprints from the reference ligands to the fingerprint of the screening compound is then estimated using the previously developed 1-NN model. The protein target of this most similar reference ligand can thus be predicted and the respective probability of the interaction actually occurring be calculated using the Bayes theorem. This probability of the prediction is displayed in the HitPick output as a score, from which the predicted targets are ranked accordingly.

3.5 *admetSAR*

The tool *admetSAR* (<http://lmm.d.ecust.edu.cn/admetSar2/>) is yet another target prediction tool based on 2D ligand-based similarity. This tool is available free of charge as a web server and was initiated by Cheng et al., Laboratory of Molecular Modeling and Design at the East China University in Shanghai [140]. Recently, in August 2018, *admetSAR* was updated and extended [141]. The presently described study was conducted before this date, so that the data presented herein were acquired with the first version of *admetSAR*. This was developed, as the name implies, as an *in silico* absorption, distribution, metabolism, excretion, toxicity (ADMET) prediction tool, including the prediction of hERG binding. The user can seek a large variety of predictions from these categories, like, e.g., being a substrate of various cytochromes, P-glycoprotein, but also an estimation of log S or permeability in Caco cells. *admetSAR* uses a collection of SVM models (described in Sect. 2.1.3) that are either classifiers (22 models) or regression models (five models). While the SVM classifiers assign screening compounds to binary classes (like, e.g., substrate of enzyme: “yes”/“no,” inhibitor of enzyme: “yes”/“no,” Caco permeable: “yes”/“no,” etc.), the SVM regression models predict continuous values (like, e.g., log S or LD $_{50}$ in rats). All models were built from an in-house database that was created manually and curated by retrieving high-quality scientific literature. This database is accessible as well free of charge (<http://lmm.d.ecust.edu.cn/admetSar1/>). All compounds are stored as MACCS keys, as implemented in OpenBabel (version 2.3.1), which are further used as inputs for the various SVM models [136].

In the initial version of *admetSAR*, two hERG models were available. The first model contained weak and strong inhibitors of hERG, while the second model contained inhibitors and non-inhibitors of hERG. Since the scope of this study was to evaluate the discrimination capacity of *admetSAR* between binders and non-binders of hERG, only the second model was considered for further use.

3.6 PASSonline

PASSonline (<http://www.way2drug.com/PASSOnline/>) is the web-accessible form of the PASS program, target prediction software developed by Poroikov and colleagues at the Institute of Biomedical Chemistry, Moscow, Russia [142, 143]. PASSonline is available free of charge. The web server operates in a ligand-based manner, by comparing screening compounds to compound datasets annotated with the respective biological activity using 2D similarity. Screening compounds are then compared to training compounds of various protein targets or disease conditions by Bayesian estimates (as outlined in Sect. 2.1.3).

As a first step, the user has to input the screening compound as a SMILES string, an sd-file, or a mol2-file. PASSonline converts the input molecule to a chemical fingerprint. Throughout PASSonline, an in-house-developed second-level MNA fingerprint is used—a variant of circular fingerprints [57]. In the second step, PASSonline compares the second-level MNA to all compounds to the training set present in PASSonline. For PASSonline, this training dataset was derived from the literature, describing either protein targets (e.g., “MAP kinase 14 inhibitor”), disease conditions (e.g., “cystic fibrosis treatment”), or simply categories assigned to compounds regarding their biological activity (e.g., “antioxidant” or “weight loss”). The latter two prediction categories somehow distinguish PASSonline from the other approaches presented in this study. Online available web servers like SEA, SuperPred, SwissTargetPrediction, HitPick, Pred-hERG, and VirtualToxLab predict discrete macromolecular targets, while PASSonline can predict also more diffuse, functional biological activities. Initially, all the activity data were derived from the MDL Drug Data Report, covering 190,000 compounds with activities assigned for over 120 different categories. When this study was carried out (March 2017), the training set of PASSonline contained 4099 categories with a total of 11,39,257 active compounds. Note that many of the active compounds in this database are present in several categories, so PASSonline does not contain 11,39,257 single compounds. A summary of the dataset is available at the PASSonline homepage (<http://www.way2drug.com/PASSOnline/methods.php>). The actual comparison of the screening compound to a compound of the dataset is made by the PASSonline algorithm. The PASSonline algorithm is a modified naïve Bayesian estimate that is used to compute both the probability of the screening compound as being active (P_a) and its probability to be inactive (P_i). A detailed description of the mathematical approach used was described elsewhere [142]. Finally, the output generated by PASSonline is a list containing the predicted activity, P_a values, and P_i values. The list is sorted in descending order of $P_a - P_i$, predicted activities where the probability of being active is high, while the probability to be inactive is low. The user is thus able to adjust the selection criteria based on the output, by not just depending on $P_a - P_i$, and hence vary the overall sensitivity and specificity.

The program PASSonline contains four training sets of different sizes for hERG that can be predicted (see Table 1).

Table 1 Training sets available for hERG in PASSonline

Activity type	Active compounds
Ether-a-go-go potassium channel 1 blocker	21
Ether-a-go-go potassium channel blocker	21
hERG 1 channel blocker	38
hERG channel blocker	1207

3.7 *Pred-hERG*

Another open-source hERG prediction platform is Pred-hERG (<http://labmol.com.br/predherg/>). The platform Pred-hERG was developed by the LabMol—Laboratory for Molecular Modeling and Drug Design at the Faculty of Pharmacy, Federal University of Goias [144, 145]. Pred-hERG operates fully ligand-based and can be classified as a 2D similarity machine-learning approach. The training data were retrieved from ChEMBL and accordingly processed. The curated dataset consists of 5984 compounds including 2191 non-blockers (activity $\geq 10 \mu M$), 2565 weak/moderate blockers ($1 \mu M \leq \text{activity} \leq 10 \mu M$), and 1228 strong blockers ($\leq 1 \mu M$). Pred-hERG computed Morgan fingerprints using the RDKit (<http://rdkit.org>) as implemented in the Konstanz Information Miner (KNIME, KNIME AG, Zurich, Switzerland). Moreover, molecular descriptors as implemented in the Chemistry Development Kit (CDK) in KNIME were computed. The dataset has been used to train MLAs based on SVMs, one as a binary model and one as a multiclass model. Both of the models were validated theoretically using a fivefold cross validation, yielding *Acc* values from 0.83 to 0.84. Use of the web server is fairly easy and can be done by submitting a molecule as SMILES string into the web form. The output produced by Pred-hERG consists of three parts: first, the binary model outputs one of the two classes “blocker” or “non-blocker,” with the respective probability. Second, the multiclass model classifies the screening compound into “strong blocker,” “weak blocker,” or “non-blocker,” again with the respective probability. Third, Pred-hERG outputs a two-dimensional structure of the screened molecule and the respective predicted probability maps (PPM) mapped onto it. In the PPMs, fragments or substructures highlighted in green contribute to hERG blockage, while those highlighted in pink counteract hERG blockage. Fragments marked with a gray color do not contribute to hERG blockage.

3.8 *VirtualToxLab*TM

VirtualToxLabTM is commercial software free of charge for academic users and was developed by A. Vedani, M. Dobler, and M. Smiesko at the Department of Pharmaceutical Sciences, University of Basel and the Foundation Biographics Laboratory 3R, Basel, Switzerland (<http://www.biograf.ch/index.php?id=projects&subid=virtualtoxlab>) [146–148]. VirtualToxLabTM is the only structure-based

method that was used during the present study. It combines flexible docking, binding affinity calculation, and linear regression into a single protocol. VirtualToxLab™ currently features 16 proteins, on which the test compounds are sampled and that are known to trigger concerning effects toxicologically *in vivo*. One of those proteins is the hERG channel. At the time when VirtualToxLab™ was developed, no experimentally solved three-dimensional model of the human hERG protein was available. Thus, the developers of VirtualToxLab™ used the homology model (as described in Sect. 2.1.5.1) of the hERG channel that was published previously by Farid et al. [109].

For VirtualToxLab™, as a first step, the compounds intended for screening need to be prepared, as described previously in Sect. 2.1.6. This can be done by either preparing the desired compounds with any external software computing three-dimensional information or with the built-in VTL builder. Next, a flexible docking protocol docks the prepared ligands into the binding site of the protein. Typical docking algorithms perform so-called rigid docking, meaning that the protein is considered as a rigid body. Flexible docking as in VirtualToxLab™ allows for the simulation of induced-fit binding. To account for such binding modes, the docking algorithm needs to keep parts of the protein, e.g., a predefined radius around the ligand binding site, flexible. This requires more computational effort, leading to longer calculation times. In VirtualToxLab™, the flexible docking is carried out by the software Cheetah [146]. Moreover, VirtualToxLab™ simulates dynamic solvation effects of the ligand and the binding pocket [147]. Finally, the flexible docking protocol outputs up to 25 poses per input ligand. For each input ligand is thus output as conformational ensemble (various poses), generated within the binding site of the protein, augmented by solvation effects and induced-fit (4D-dataset). The conformational ensemble of each input ligand is then used to calculate binding affinity to the respective protein, using 4D Boltzmann scoring in the software BzScore4D [148]. BzScore4D samples each conformational ensemble coming from the flexible docking protocol using Monte Carlo simulation with Metropolis criterion, first in a water box and second in the protein's binding site. Each functional group is then scored individually in both states and the change in free energy calculated. Finally, VirtualToxLab™ predicts the toxic potential for each compound to the respective target. In the case of hERG, this corresponds to the degree of interference with the channel.

4 Results and Discussion

4.1 Study Setup

The aim of this case study was to evaluate whether publicly available tools are able to predict hERG-modulated toxicity. In the last few decades, hERG has gained much research attention as an antitarget. Consequently, a great deal of research has been done on hERG, especially in the field of target prediction. Being able to predict

hERG efficiently using *in silico* techniques, representing low costs and time commitments, has thus represented a constant objective for medicinal chemists. Surprisingly, comparatively little effort has been placed on the development of robust hERG screening tools (*in vitro* and *in silico*) for natural products. Many natural products have been characterized as hERG blockers, even some of those frequently used in phytopharmaceuticals [149]. Thus, our focus was set on evaluating the predictive power of some open-access hERG prediction tools toward natural products. To do so, eight open-access web servers (presented in Sect. 3) with two pharmacophore models/pharmacophore model ensembles (as published by Kratz et al. [96]) and a dataset of 278 compounds were screened (see Sect. 4.2).

4.2 Dataset Used in This Study

The dataset gathered for this study contained a total of 277 compounds. The dataset could be subdivided into 188 natural products and 90 synthetic compounds. The natural products subset could be further subdivided based on their origin from the alkaloid compound class (129 compounds) and non-alkaloid compound class (59 compounds). Of the alkaloids, 37 compounds were categorized as “strong blockers,” 28 as “moderate blockers,” and 64 as “non-blockers.” From the non-alkaloids, seven compounds were categorized as “strong blockers,” six as “moderate blockers,” and 46 as “non-blockers.” Among the synthetic compounds, 57 were categorized as “blockers” and 33 as “non-blockers.” The dataset and the respective subsets are visualized in Fig. 11. The natural products were derived from a

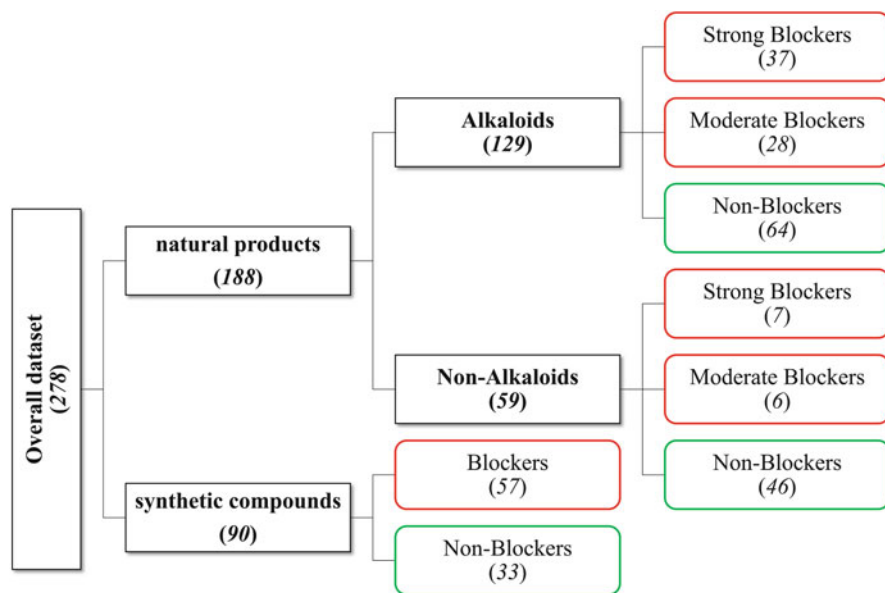


Fig. 11 Dataset and the respective subsets visualized. Single-colored red rectangles indicate compounds that block the hERG channel; single-colored green fields indicate compounds that do not block the hERG channel

comprehensive review on natural product hERG blockers, authored by Kratz et al. [149]. The synthetic compounds were used from a study conducted by Kratz et al. and our own department in 2014 [96].

4.3 Data Curation and Pre-processing

To be able to compare the respective hits in a semi-automated and efficient way, the target nomenclature for hERG had to be normalized. Moreover, some of the tools used contain different models with each different names referring to hERG. The SMILES strings for all structures were generated by using the “copy as → SMILES” function in ChemDraw Professional (Perkin Elmer, Waltham, MA, USA) and pasting them into an Excel sheet.

The exact data curation for each tool is described as follows: SEA uses the ChEMBL/UniProt nomenclature, and thus outputs the full target name “Potassium voltage-gated channel subfamily H member 2,” as well as other metadata. Only predictions containing the key “Potassium voltage-gated channel subfamily H member 2” in the “Description” column were considered as predicted actives. Compounds without this prediction were considered as predicted inactive. In SuperPred, SMILES strings were imported and the “Target Prediction” option selected. SuperPred uses the nomenclature derived from ChEMBL. Only predictions containing the key “KCNH2” in the “Name” column were considered as predicted actives. Compounds without this prediction were considered to be predicted inactive. In SwissTargetPrediction, SMILES strings were posted to the server. SwissTargetPrediction uses the nomenclature derived from ChEMBL. Only predictions containing the key “Potassium voltage-gated channel subfamily H member 2” in the “Target” column were considered as predicted actives. Compounds lacking this annotation were considered as predicted inactive. In HitPick, SMILES strings were posted to the server, using the “Target Prediction” suite. HitPick outputs various metadata on each predicted interaction, including the gene names of the respective proteins. Only predictions containing the key “KCNH2” in the “Target” column were considered as predicted actives. Compounds without this annotation were considered as predicted inactive. In admetSAR, SMILES strings were posted to the server, using the “predict” suite. The admetSAR contains two hERG models: T_hERG_I, classified into “Weak inhibitor” and “Strong inhibitor,” and the model “T_hERG_II” classified into “Inhibitor” and “Non-inhibitor.” Only the model “T_hERG_II” was used, since the aim of this study was to evaluate the predictive power of such tools between true actives and true inactives. Compounds predicted by “T_hERG_II” as “Inhibitor” were considered as predicted active, while those predicted as ‘Non-inhibitors’ were considered as predicted inactive. In PASSonline, SMILES strings were uploaded to the server. As outlines in Sect. 3.6, Table 1, PASSonline contains four models for hERG. In this study, all of the four were considered, since all of them were binary classifiers trained on the discrimination between true actives and inactives. Thus, all compounds that were classified into one

of those four classes by PASSonline were considered as predicted actives and all remaining compounds as predicted inactive. Moreover, PASSonline's probability output is composed of P_a and P_i values. We subtracted P_i from P_a , to calculate an overall probability and ranked the prediction according to this information. Large differences in P_a and P_i may then be considered as more probable than small differences. In Pred-hERG, SMILES strings were posted to the server. Similar to admetSAR, Pred-hERG contains two models: one binary classifier and one multiclass classifier. Again, only the binary classifier was evaluated, since the scope of this study was limited to this instance. Accordingly, the prediction "blocker" given in the row "Binary Pred" was considered as predicted active, while the prediction "non-blocker" was considered as predicted inactive. In VirtualToxLab™, each compound was submitted as single sd-file. Finally, the binding affinity in the "hERG" column was used. According to VirtualToxLab™, binding affinities smaller than 100 μM were considered as predicted active, while binding affinities greater than 100 μM were considered as predicted inactive. For the in-house pharmacophore model collection, the compounds were prepared into one single sd-file, and preprocessed with idbgen application in LigandScout 3.03b (Inte: Ligand, Vienna, Austria), using "best" settings [150]. Screening was performed in LigandScout 3.03b, using the six pharmacophore models described by Kratz et al. [96] in cooperative mode. Compounds that mapped to one or more pharmacophore models were considered as predicted actives, and the remaining compounds were considered as predicted inactive. For the Discovery Studio pharmacophore models, the same single sd-file of the compounds was used as input. The ligands were prepared using the "build 3D database" application and the screening conducted using the "search 3D database" application, both integrated in Discovery Studio version 4.5 (Dassault Systèmes, Vélizy-Villacoublay, France). The pharmacophore model used in screening was described by Kratz et al. [96], and referenced there as "Catalyst model."

4.4 Post-processing of Data

The data were collected by inserting SMILES strings, or sd-files, if required, to the respective tool. The resulting output was gathered in Excel spreadsheets, csv-files, or sd-files.

Every approach used during this study outputs the respective results in different formats. To be able to evaluate the results obtained from both online accessible target prediction tools and our in-house pharmacophore models, the data had to be processed and homogenized first. A custom-built KNIME workflow was created to fulfill this task (illustrated in Fig. 12). The KNIME workflow has ten inputs: one for the in vitro results and nine for the nine different in silico tools (represented by the single-colored rectangles). Moreover, the workflow produces one single output. The resulting tables could be saved automatically as csv-files (yellow-framed rectangle). The output is subdivided three times, accounting for the overall dataset, natural

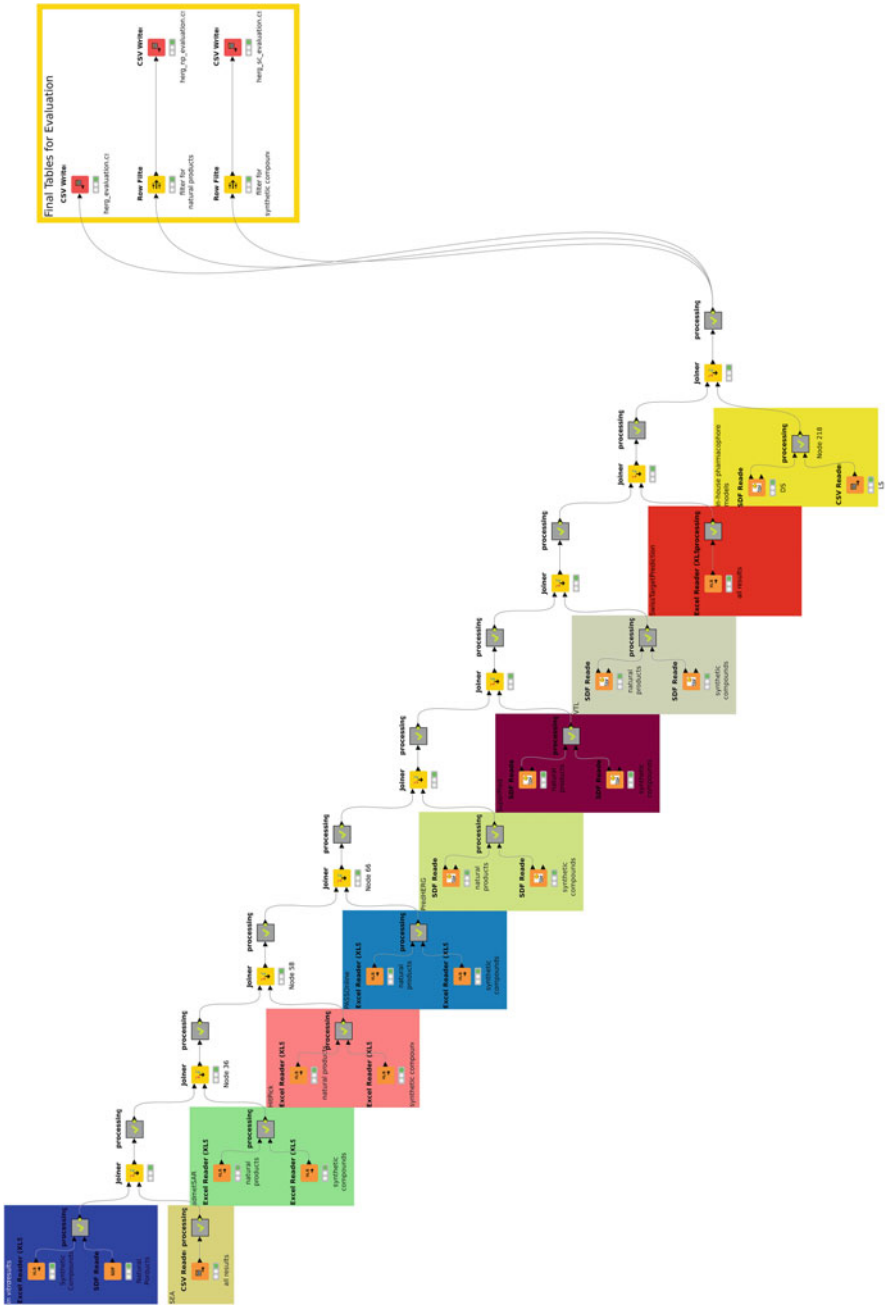


Fig. 12 KNIME workflow used to process data

products, and synthetic compounds. The complete table obtained from the KNIME workflow described above was imported subsequently into Jupyter Notebook (Project Jupyter, <https://jupyter.org/>) to calculate quantitative performance metrics for each tool and to compute ROC curves.

4.4.1 Performance with Natural Products

The ROC curves produced for the natural products subset are given in Fig. 13, and their quantitative performance metrics for the natural products subset are provided in Table 2. In general, all of the used tools performed poorly, with the exception of Pred-hERG, which showed an intermediate performance. Pred-hERG, showing the best predictive power according to ROC-AUC (0.63), produced an overall *Acc* of 0.66, while being more specific (0.90) than sensitive (0.33), which stands for the more accurate classification of true negatives as such. The *Pr*, on the other hand, was intermediate (0.70), and thus the *F*₁ score relatively low (0.45). The admetSAR *F*₁ score (0.46) was slightly better than the one produced by Pred-hERG, proving an improved sensitivity/precision ratio. This might seem confusing, since the other computed metrics indicate poor performance (*Acc*, *Se*, *Sp*, *Pr*, and ROC-AUC). This occurred because the *F*₁ score represents a harmonic mean of the similarly poor *Se* and *Pr*. For the poor *F*₁ score of Pred-hERG, the good *Pr* was made worse by the poor *Se*. Interestingly, in admetSAR, the ROC curve shows that especially high-

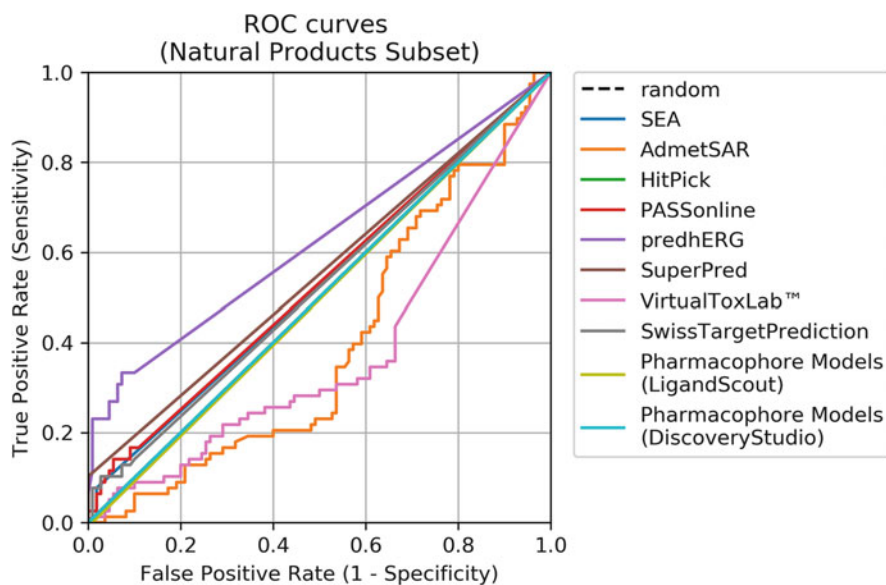


Fig. 13 ROC curves of the tools used, describing their predictive power toward natural products investigated in this study

Table 2 Quantitative performance metrics of the used tools toward natural products

Tool	TN	FP	FN	TP	<i>Acc</i>	<i>Se</i>	<i>Sp</i>	<i>Pr</i>	F_1	ROC-AUC
SEA	108	2	72	6	0.61	0.08	0.98	0.75	0.14	0.53
admetSAR	73	37	44	34	0.57	0.44	0.66	0.48	0.46	0.39
HitPick	110	0	78	0	0.59	0.00	1.00	0.00	0.00	0.50
PASSonline	98	12	65	13	0.59	0.17	0.89	0.52	0.25	0.53
Pred-hERG	99	11	52	26	0.66	0.33	0.90	0.70	0.45	0.63
SuperPred	110	0	70	8	0.63	0.10	1.00	1.00	0.19	0.55
VirtualToxLab	37	73	44	34	0.38	0.44	0.34	0.32	0.37	0.38
SwissTargetPrediction	99	11	67	11	0.59	0.14	0.90	0.50	0.22	0.52
Pharmacophore model (LigandScout)	109	1	78	0	0.58	0.00	0.99	0.00	0.00	0.50
Pharmacophore model (Discovery Studio)	110	0	78	0	0.59	0.00	1.00	0.00	0.00	0.50

ranked predictions (bottom left, $\text{Slope}_{\text{ROC}} < 1$) are more often incorrect than correct. Predictions below the median of the probability become more accurate ($\text{Slope}_{\text{ROC}} > 1$). Overall, the admetSAR ROC curve performed worse than random for the natural products subset (ROC-AUC = 0.39), but the accuracy was slightly better than random (0.57). VirtualToxLab™ also produced an ROC-AUC worse than random (0.38), and more than half of the predictions were incorrect ($\text{Acc} = 0.38$) for the natural products subset. Accordingly, the remaining metrics (Se , Sp , Pr , F_1) also show a performance worse than random. The two in-house pharmacophore approaches (LigandScout and Discovery Studio models) showed a perfectly random performance according to ROC-AUC (both 0.50), but more than half of the predictions made were correct ($\text{Acc} = 0.58$ and 0.59, respectively). Both models appeared very good at detecting true negatives as such ($\text{Sp} = 0.99$ and 1.00, respectively), but performed poorly in detecting true binders ($\text{Se} = 0$ both). Accordingly, Pr and F_1 were also calculated as zero. The same was true for HitPick, which showed a random ROC-AUC (0.50), a Se of zero, and a Sp of one. The overall accuracy was the same as for both pharmacophore models better than random ($\text{Acc} = 0.59$). SEA, SwissTargetPrediction, and PASSonline showed very similar performance, slightly superior to random (ROC-AUC = 0.53, 0.53, 0.52, respectively). Similar to the pharmacophore models and HitPick, Se was very poor for all of the three (0.08 to 0.17), while Sp was very good (0.89 to 0.98). SEA showed the best Sp (0.98), PASSonline showed the best F_1 (0.25), and SwissTargetPrediction showed the best Pr (0.75). Acc was again slightly superior to the group of the two pharmacophore models and HitPick (0.59–0.61). A slightly better performance was achieved by SuperPred. SuperPred showed the second-best ROC-AUC (0.55) while posing 63% of the predictions correct ($\text{Acc} = 0.63$). As seen with most of the other models, Se was very low (0.10), while Sp was perfect (1.00). SuperPred even achieved the highest possible Pr (1.00).

It remains to be said, that according to the quantitative metrics obtained, Pred-hERG was clearly superior, followed by SuperPred. A clear trend could be observed

with all of the used tools, namely, the poor Se values and high Sp values. These properties could be beneficial for certain applications in drug discovery, but for that considered here in being of toxicological interest, the exact opposite would be desirable. Such applications should be greatly more sensitive than specific. For example, high Sp , Pr , and F_1 values are preferable in classical VS, because this means that the hit list will be very much enriched with true actives. In toxicologically relevant VS, it would be better to detect possibly all potentially harmful compounds. This property is indicated by high Se values, even if Pr is low. This would mean that most of the true hERG binders are detected by this model, even if it produces many FPs. With this being said, admetSAR produced the best results, followed by VirtualToxLab™, even though the performances of both tools were still comparably poor.

4.4.2 Performance with Synthetic Compounds

The ROC curves produced for the synthetic compounds subset are provided in Fig. 14, and the quantitative performance metrics for the synthetic compounds subset are provided in Table 3. For the synthetic compound subset, the overall performance of all tools was more satisfactory than for the natural products subset. All tools performed intermediate to good, with the exception of the Discovery Studio

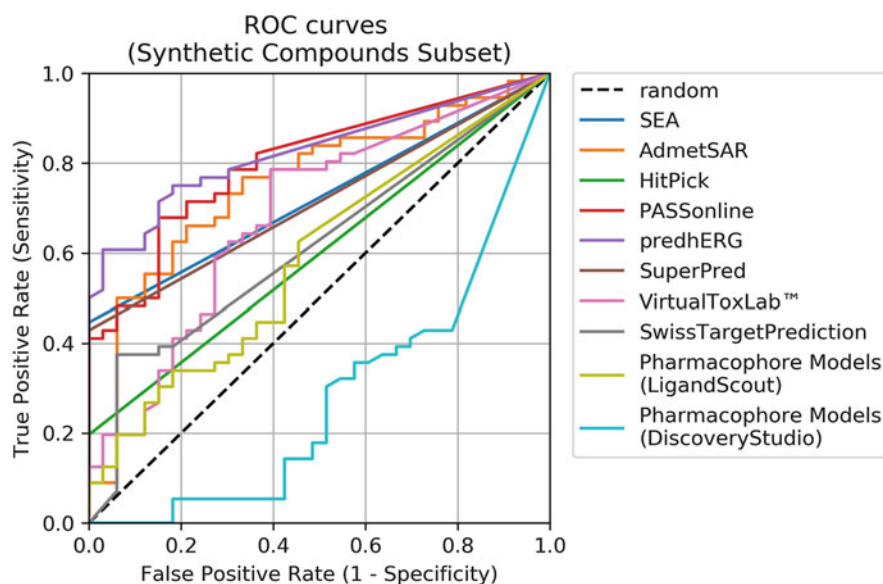


Fig. 14 ROC curves of the tools used, describing their predictive power toward synthetic compounds investigated in this study

Table 3 Quantitative performance metrics of the used tools toward synthetic compounds

Tool	TN	FP	FN	TP	<i>Acc</i>	<i>Se</i>	<i>Sp</i>	<i>Pr</i>	F_1	ROC-AUC
SEA	33	0	31	25	0.65	0.45	1.00	1.00	0.62	0.72
admetSAR	17	16	8	48	0.73	0.86	0.52	0.75	0.80	0.76
HitPick	33	0	45	11	0.49	0.20	1.00	1.00	0.33	0.60
PASSonline	21	12	10	46	0.75	0.82	0.64	0.79	0.81	0.80
Pred-hERG	23	10	12	44	0.75	0.79	0.70	0.81	0.80	0.83
SuperPred	33	0	32	24	0.64	0.43	1.00	1.00	0.60	0.71
VirtualToxLab	14	19	10	46	0.67	0.82	0.42	0.71	0.76	0.69
SwissTargetPrediction	27	6	34	22	0.55	0.39	0.82	0.79	0.52	0.62
Pharmacophore model (LigandScout)	18	15	20	36	0.61	0.64	0.55	0.71	0.67	0.61
Pharmacophore model (Discovery Studio)	7	26	32	24	0.35	0.43	0.21	0.48	0.45	0.28

pharmacophore model. For the Discovery Studio pharmacophore model, the metrics indicated a performance worse than random for this test set. In turn, HitPick, SwissTargetPrediction, SEA, and SuperPred showed similar performance. HitPick shows the typical behavior observed with the natural products subset, and the model is way more specific (1.00) than sensitive (0.20). The *Pr* is very good (1.00), and accordingly F_1 improved as a consequence (0.33). SwissTargetPrediction also shows the tendency of being more specific (0.82) than sensitive (0.39), even though the discrepancy was not as large as with HitPick. The *Pr* is good (0.79) and the F_1 moderate (0.52). Also, SEA and SuperPred performed similarly, but as seen already with SwissTargetPrediction, the discrepancy between *Se* and *Sp* was smaller than observed with HitPick. Moreover, they produced quite distinct ROC-AUCs (0.72, 0.60, 0.71, 0.62, respectively) and *Acc* (0.65, 0.49, 0.64., 0.55, respectively). Another group of tools also showed very similar performances to one another: admetSAR, PASSonline, Pred-hERG, VirtualToxLab™, and the LigandScout pharmacophore models. They all have in common that their *Se/Sp* ratio is comparatively good. From this group, the LigandScout pharmacophore model's performance is worse than that of the other tools. admetSAR, PASSonline, Pred-hERG, and VirtualToxLab™ showed both good *Se* and good *Pr* scores. Accordingly, they all have very good F_1 scores. Moreover, in this group, admetSAR, PASSonline, Pred-hERG, and VirtualToxLab™ show a good early enrichment in ROC curves (by rising steeply at the left half of the plot, and decreasing their slope also at the left half of the plot). This indicated that especially highly ranked predictions (high probability score assigned by the respective tool) show very favorable ratios between TPR and FPR. The LigandScout pharmacophore models also show this property to a less extent. Here, the slope of the ROC curve was evenly distributed along the whole plot.

According to the quantitative performance metrics, Pred-hERG and PASSonline showed the best performance in predicting synthetic hERG binders. As described in

Sect. 4.4.1, high *Se* values are desirable for a VS that is toxicologically relevant. This being said, admetSAR showed again the best performance, followed by PASSonline, VirtualToxLab™, and Pred-hERG. The LigandScout pharmacophore models showed a good tendency, but were unable to retrieve as many true actives as admetSAR, PASSonline, Pred-hERG, and VirtualToxLab™. In this context, the performances of these tools can be described as good, while the LigandScout pharmacophore model was moderate; HitPick, SwissTargetPrediction, SEA, and SuperPred performed poor; and the Discovery Studio pharmacophore model showed poor performance.

4.4.3 Overall Performance

The ROC curves produced for the complete dataset are provided in Fig. 15, and the quantitative performance metrics for the synthetic compounds subset is provided in Table 4. As discussed in Sect. 4.4.2, the predictive accuracy was good for the synthetic compounds subset and poor for the natural products subset, resulting in a moderate performance for the complete dataset. According to the ROC curves, Pred-hERG showed the best performance, followed by PASSonline, SuperPred, and SEA. Moreover, they all show good early enrichment in the ROC curve. VirtualToxLab™ also showed a good early enrichment but failed to maintain this TPR/FPR ratio along the plot. Therefore, the highly ranked predictions posed by VirtualToxLab™ are quite reliable, but intermediate-to-low probabilities tended to be more error-prone.

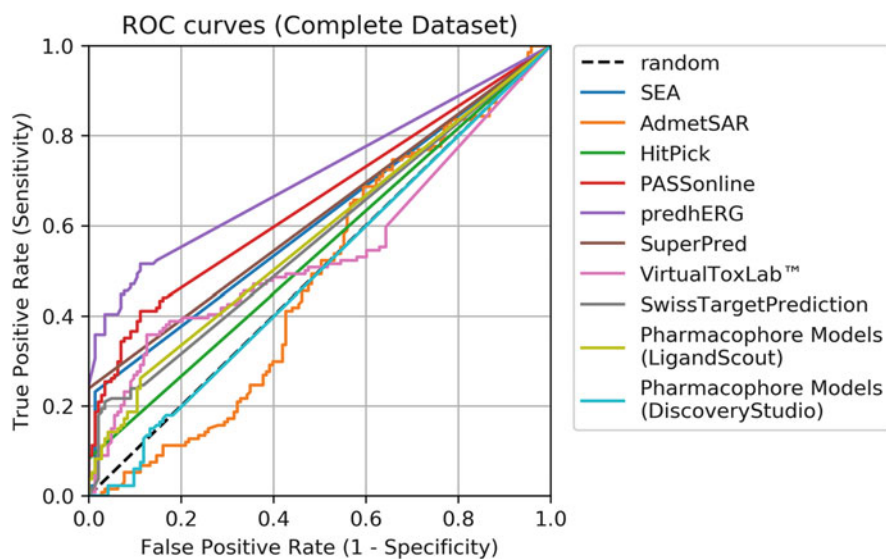


Fig. 15 ROC curves of the tools used, describing their predictive power toward synthetic compounds investigated in this study

Table 4 Quantitative performance metrics of the used tools toward the complete dataset

Tool	TN	FP	FN	TP	Acc	Se	Sp	Pr	F_1	ROC-AUC
SEA	141	2	103	31	0.62	0.23	0.99	0.94	0.37	0.61
admetSAR	90	53	52	82	0.62	0.61	0.63	0.61	0.61	0.48
HitPick	143	0	123	11	0.56	0.08	1.00	1.00	0.15	0.54
PASSonline	119	24	75	59	0.64	0.44	0.83	0.71	0.54	0.65
Pred-hERG	122	21	64	70	0.69	0.52	0.85	0.77	0.62	0.71
SuperPred	143	0	102	32	0.63	0.24	1.00	1.00	0.39	0.62
VirtualToxLab	51	92	54	80	0.47	0.60	0.36	0.47	0.52	0.54
SwissTargetPrediction	126	17	101	33	0.57	0.25	0.88	0.66	0.36	0.57
Pharmacophore models (LigandScout)	127	16	98	36	0.59	0.27	0.89	0.69	0.39	0.58
Pharmacophore models (Discovery Studio)	117	26	110	24	0.51	0.18	0.82	0.48	0.26	0.49

The ROC curve of SuperPred, in comparison to the other three tools in the latter group, rose very steeply and then got connected with a linear line to the top right corner of the plot. This indicated that the highest probabilities calculated by SuperPred were always correct, although it should be noted that such probabilities were assigned to only a very few compounds (those predicted as hERG inhibitors), which were in fact all correct (32 TPs, 0 FPs). The remaining 245 compounds did not produce a prediction for hERG, for which, in these cases, was set to a probability of zero. This produced of course the highest reachable precision of 1.00. The same phenomenon can be observed with SEA. The SEA predicted 31 compounds correctly as hERG inhibitors (TP) and two were falsely classified as such (FP). This gave an ROC curve very similar to that of SuperPred, with very similar ROC-AUCs (0.61 and 0.62, respectively). Thus, the resulting Pr produced by SEA is bit lower than that of SuperPred, but still very good (0.94). The next group of tools producing similar ROC-AUCs comprised SwissTargetPrediction, HitPick, and the LigandScout pharmacophore models. The ROC curve produced by HitPick was very similar to those produced by SuperPred and SEA. Accordingly, the same interpretation is true also for HitPick. The main difference from the performance of SuperPred and SEA is that HitPick showed a much lower Se (0.08), compared to 0.24 and 0.23, respectively. As a consequence, the resulting F_1 score (0.15) was also much lower than those of SuperPred (0.39) and SEA (0.37). The values of Acc , Sp , Pr , and ROC-AUC were comparable. SwissTargetPrediction and the LigandScout pharmacophore models performed in a similar manner in all calculated metrics, with the LigandScout pharmacophore model always performing slightly better. The main difference between the performances of those two tools was that SwissTargetPrediction showed an early enrichment in the ROC curve, while the LigandScout pharmacophore models did not. The admetSAR gives an ROC curve showing a performance below random, resulting in an ROC-AUC of 0.48. On the other hand, SEA was the overall best (0.61), with an acceptable Pr (0.61). Accordingly, the F_1 score was also 0.61, and was only outperformed by Pred-hERG (0.62). Moreover, the inverse behavior of the ROC curves toward VirtualToxLab™ could be seen. While VirtualToxLab™ showed an early enrichment, which depletes over time, admetSAR seemed to produce more reliable predictions when the calculated probabilities were intermediate to low. The Discovery Studio pharmacophore model's performance was poor and showed random accuracy for the natural products subset, and an accuracy below random for the synthetic compounds subset.

In summary, the overall performance of all tools taken together was disappointing. For synthetic compounds, there are definitely free-of-charge online tools available that are easy to use and can give reliable results. For natural products, there seems to be a great need for improvement. This might be due to the fact that none of the tools employed was explicitly designed for use with natural products. It is well known that natural products differ from synthetic compounds, which was discussed extensively in several publications [29].

5 Conclusion and Outlook

The predictive power of open-access activity prediction tools was assessed. All of the tools evaluated in this study performed more effectively for synthetic compounds than for natural products. This outcome can be mainly addressed to the distinct chemical space occupied by natural products when compared with synthetic compounds. Another factor is that the training sets used for the available tools are always composed of bioactivity data derived from the scientific literature. Regarding hERG, this literature is mainly comprised by synthetic compounds, and less by natural products. Moreover, the overall awareness of hERG-blocking properties of natural products is much lower when compared to synthetic compounds. Moreover, the models used are able to handle synthetic compounds better than they do natural products.

Pharmacophore-based approaches and VirtualToxLab™ were mostly found to be outperformed by the alternative approaches used. These pharmacophore-based approaches were generated out of just a few training compounds, which are representative for the chemical space of hERG blockers. The resulting models then screen for compounds that can exert the same hypothetical binding modes. VirtualToxLab™ does not use training compounds at all, since it relies fully on molecular docking. Molecular docking requires to define one single binding site, in which the poses are sampled. As known today, hERG blockers bind to several binding sites of the protein, which naturally produces several possible binding modes. This may have a significant impact on the performance of these two approaches. In comparison, the other approaches presented using two-dimensional ligand-based similarity comparisons have a much different data foundation. They rely on thousands of training compounds, against which the compounds screened are compared. Such approaches are largely independent of binding sites, and perform best if the training sets are large.

In silico prediction methodology for the hERG blocking potential of natural products still has to mature. To date, few efforts have been made to design in silico prediction tools for hERG, specifically for natural products, or to explicitly include natural products. Nevertheless, substantial progress should be made in this field over the next few decades. Wang and MacKinnon have opened the door for structure-based modeling on hERG, and to a better structural understanding of hERG [10]. Furthermore, the knowledge of both synthetic compound and natural product hERG blockers is constantly increasing. This should ultimately lead to more comprehensive and more accurate in silico tools for this particular target, for both synthetic compounds and for compounds biosynthesized by Nature.

Acknowledgments The research has been funded by GECT Euregio Tirol-Südtirol-Trentino (IPN55). The authors thank OpenEye Scientific Software and Inte:Ligand for their academic licenses. Daniela Schuster is an Ingeborg Hochmair Professor at the University of Innsbruck.

References

1. Vandenberg JI, Perry MD, Perrin MJ, Mann SA, Ke Y, Hill AP (2012) hERG K⁺ channels: structure, function, and clinical significance. *Physiol Rev* 92:1393
2. Sanguinetti MC, Tristani-Firouzi M (2006) hERG potassium channels and cardiac arrhythmia. *Nature* 440:463
3. Haverkamp W, Breithardt G, Camm AJ, Janse MJ, Rosen MR, Antzelevitch C, Escande D, Franz M, Malik M, Moss A, Shah R (2000) The potential for QT prolongation and proarrhythmia by non-antiarrhythmic drugs: clinical and regulatory implications. Report on a policy conference of the European Society of Cardiology. *Eur Heart J* 21:1216
4. Siramshetty VB, Nickel J, Omieczynski C, Gohlke B-O, Drwal MN, Preissner R (2016) WITHDRAWN—a resource for withdrawn and discontinued drugs. *Nucleic Acids Res* 44: D1080
5. The non-clinical evaluation of the potential for delayed ventricular repolarization (QT interval prolongation), by human pharmaceuticals, S7B (Step 5) (2005). International conference on harmonization (ICH) of technical requirements for registration of pharmaceuticals for human use. <https://www.ich.org>
6. Clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs, E14 (Step 5) (2005). International conference on harmonization (ICH) of technical requirements for registration of pharmaceuticals for human use. <https://www.ich.org>
7. Chi KR (2013) Revolution dawning in cardiotoxicity testing. *Nature Rev Drug Discov* 12:565
8. Mitcheson JS, Chen J, Lin M, Culberson C, Sanguinetti MC (2000) A structural basis for drug-induced long QT syndrome. *Proc Natl Acad Sci U S A* 97:12329
9. Kalyanamoorthy S, Barakat KH (2018) Development of safe drugs: the hERG challenge. *Med Res Rev* 38:525
10. Wang W, MacKinnon R (2017) Cryo-EM structure of the open human ether-à-go-go-related K⁺ channel hERG. *Cell* 169:422
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235
12. Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2014) The Protein Data Bank archive as an open data resource. *J Comput Aided Mol Des* 28:1009
13. Burley SK, Berman HM, Christie C, Duarte JM, Feng Z, Westbrook J, Young J, Zardecki C (2018) RCSB Protein Data Bank: sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci* 27:316
14. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, Christie C, Dalenberg K, Duarte JM, Dutta S, Feng Z, Ghosh S, Goodsell DS, Green RK, Guranović V, Guzenko D, Hudson BP, Kalro T, Liang Y, Lowe R, Namkoong H, Peisach E, Periskova I, Prlić A, Randle C, Rose A, Rose P, Sala R, Sekharan M, Shao C, Tan L, Tao Y-P, Valasatava Y, Voigt M, Westbrook J, Woo J, Yang H, Young J, Zhuravleva M, Zardecki C (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* 47:D464
15. Lawson CL, Baker ML, Best C, Bi C, Dougherty M, Feng P, van Ginkel G, Devkota B, Lagerstedt I, Ludtke SJ, Newman RH, Oldfield TJ, Rees I, Sahni G, Sala R, Velankar S, Warren J, Westbrook JD, Henrick K, Kleywegt GJ, Berman HM, Chiu W (2011) EMDDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res* 39:D456
16. Hosaka Y, Iwata M, Kamiya N, Yamada M, Kinoshita K, Fukunishi Y, Tsujimae K, Hibino H, Aizawa Y, Inanobe A, Nakamura H, Kurachi Y (2007) Mutational analysis of block and facilitation of hERG current by a class III anti-arrhythmic agent, nifekalant. *Channels* 1:1988
17. Perry M, Sanguinetti M, Mitcheson J (2010) Symposium review: revealing the structural basis of action of hERG potassium channel activators and blockers. *J Physiol* 588:31577

18. Saxena P, Zangerl-Plessl EM, Linder T, Windisch A, Hohaus A, Timin E, Hering S, Stary-Weinzinger A (2016) New potential binding determinant for hERG channel inhibitors. *Sci Rep* 6:24182
19. Perry M, Sachse FB, Sanguinetti MC (2007) Structural basis of action for a human ether-a-go-go-related gene 1 potassium channel activator. *Proc Natl Acad Sci U S A* 104:13827
20. Perry M, Sachse FB, Abbruzzese J, Sanguinetti MC (2009) PD-118057 contacts the pore helix of hERG1 channels to attenuate inactivation and enhance K⁺ conductance. *Proc Natl Acad Sci U S A* 106:20075
21. Yu Z, Klaasse E, Heitman LH, Ijzerman AP (2014) Allosteric modulators of the hERG K⁺ channel: Radioligand binding assays reveal allosteric characteristics of dofetilide analogs. *Toxicol Appl Pharmacol* 274:78
22. Vilums M, Overman J, Klaasse E, Scheel O, Brussee J, Ijzerman AP (2012) Understanding of molecular substructures that contribute to hERG K⁺ channel blockade: synthesis and biological evaluation of E-4031 analogues. *ChemMedChem* 7:107
23. Durdagi S, Deshpande S, Duff HJ, Noskov SY (2012) Modeling of open, closed, and open-inactivated states of the hERG1 channel: structural mechanisms of the state-dependent drug binding. *J Chem Inf Model* 52:2760
24. Xiao Y, Liang A, Li Z (2012) A comparison of the performance and application differences between manual and automated patch-clamp techniques. *Curr Chem Gen* 6:87
25. Villoutreix BO, Taboureau O (2015) Computational investigations of hERG channel blockers: new insights and current predictive models. *Adv Drug Deliv Rev* 86:72
26. Zhang C, Zhou Y, Gu SK, Wu ZR, Wu WJ, Liu CM, Wang KD, Liu GX, Li WH, Lee PW, Tang Y (2016) In silico prediction of hERG potassium channel blockage by chemical category approaches. *Toxicol Res* 5:570
27. Rodolpho CB, Vinicius MA, Meryck FBS, Eugene M, Denis F, Alexander T, Carolina HA (2014) Tuning hERG out: antitarget QSAR models for drug development. *Curr Top Med Chem* 14:1399
28. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings IPII of original article: S0169-409X(96)00423-1. The article was originally published in 1997. *Adv Drug Del Rev* 23:3
29. Feher M, Schmidt JM (2003) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J Chem Inf Comput Sci* 43:218
30. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, Karlsson A, Al-Lazikani B, Hersey A, Oprea TI, Overington JP (2016) A comprehensive map of molecular drug targets. *Nat Rev Drug Discov* 16:19
31. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40:D1100
32. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* 43:W612
33. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR (2017) The ChEMBL database in 2017. *Nucleic Acids Res* 45:D945
34. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH (2016) PubChem Substance and Compound databases. *Nucleic Acids Res* 44:D1202
35. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* 37:W623

36. Johnson MA, Maggiora GM (1990) Concepts and applications of molecular similarity. *J Comput Chem* 13. Wiley-Interscience, New York
37. Maggiora G, Vogt M, Stumpfe D, Bajorath J (2014) Molecular similarity in medicinal chemistry. *J Med Chem* 57:3186
38. Feng Y, Mitchison TJ, Bender A, Young DW, Tallarico JA (2009) Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat Rev Drug Discov* 8:567
39. Wagner BK, Clemons PA (2009) Connecting synthetic chemistry decisions to cell and genome biology using small-molecule phenotypic profiling. *Curr Opin Chem Biol* 13:539
40. Hart CP (2005) Finding the target after screening the phenotype. *Drug Discov Today* 10:513
41. Lee J, Bogyo M (2013) Target deconvolution techniques in modern phenotypic profiling. *Curr Opin Chem Biol* 17:118
42. Cavalla D, Oerton E, Bender A (2017) Drug repurposing review. In: Chackalamannil S, Rotella D, Ward S (eds) *Comprehensive medicinal chemistry III*, 3rd edn. Elsevier Science, Amsterdam, p 11
43. Jenkins JL, Bender A, Davies JW (2006) In silico target fishing: predicting biological targets from chemical structure. *Drug Discov Today Technol* 3:413
44. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Pujadas G, Garcia-Vallve S (2015) Tools for in silico target fishing. *Methods* 71:98
45. Koutsoukas A, Simms B, Kirchmair J, Bond PJ, Whitmore AV, Zimmer S, Young MP, Jenkins JL, Glick M, Glen RC, Bender A (2011) From in silico target prediction to multi-target drug design: current databases, methods and applications. *J Proteomics* 74:2554
46. Ekins S (2018) *Computational toxicology: risk assessment for chemicals*. Wiley, New York
47. University of Alberta and the Metabolomics Innovation Centre (2019) DrugBank. <https://www.drugbank.ca/>
48. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 34:D668
49. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 36:D901
50. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS (2011) DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res* 39:D1035
51. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 42:D1091
52. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucl Acids Res* 46:D1074
53. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G (2015) Molecular fingerprint similarity search in virtual screening. *Methods* 71:58
54. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Model* 42:1273
55. Daylight Chemical Information Systems, Inc. (2019) Daylight chemical information systems. <http://www.daylight.com/>
56. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50:742
57. Filimonov D, Poroikov V, Borodina Y, Glorizova T (1999) Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. *J Chem Inf Model* 39:666
58. Rogers DJ, Tanimoto TT (1960) A computer program for classifying plants. *Science* 132:1115

59. Bajusz D, Rácz A, Héberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* 7:20
60. Schuster D (2019) Fingerprints and pharmacophores. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (eds) *Encyclopedia of bioinformatics and computational biology*. Academic, Oxford, UK, p 619
61. Fischer E (1894) Einfluss der Configuration auf die Wirkung der Enzyme. *Ber Dtsch Chem Ges* 27:2985
62. Grant JA, Gallardo MA, Pickup BT (1996) A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J Comput Chem* 17:1653
63. Nickel J, Gohlke B-O, Erehman J, Banerjee P, Rong WW, Goede A, Dunkel M, Preissner R (2014) SuperPred: update on drug classification and target prediction. *Nucleic Acids Res* 42:W26
64. Rush TS, Grant JA, Mosyak L, Nicholls A (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* 48:1489
65. Hawkins PCD, Skillman AG, Nicholls A (2007) Comparison of shape-matching and docking as virtual screening tools. *J Med Chem* 50:74
66. Ballester PJ, Finn PW, Richards WG (2009) Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. *J Mol Graphics Model* 27:836
67. Kumar A, Zhang KYJ (2018) Advances in the development of shape similarity methods and their application in drug discovery. *Front Chem* 6
68. Armstrong MS, Morris GM, Finn PW, Sharma R, Richards WG (2009) Molecular similarity including chirality. *J Mol Graphics Model* 28:368
69. Armstrong SM, Morris GM, Finn PW, Sharma R, Moretti L, Cooper RI, Richards WG (2010) ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J Comput Aid Mol Des* 24:789
70. Hsin K-Y, Morgan HP, Shave SR, Hinton AC, Taylor P, Walkinshaw MD (2011) EDULISS: a small-molecule database with data-mining and pharmacophore searching capabilities. *Nucleic Acids Res* 39:D1042
71. Schreyer AM, Blundell T (2012) USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *J Cheminform* 4:27
72. Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M (2006) Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J Med Chem* 49:6802
73. Panteleev J, Gao H, Jia L (2018) Recent applications of machine learning in medicinal chemistry. *Bioorg Med Chem Lett* 28:2807
74. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today* 20:318
75. Lo Y-C, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 23:1538
76. Dubus E, Ijjaali I, Petitet F, Michel A (2006) In silico classification of hERG channel blockers: a knowledge-based strategy. *ChemMedChem* 1:622
77. Sun H (2006) An accurate and interpretable Bayesian classification model for prediction of hERG liability. *ChemMedChem* 1:315
78. Li Q, Jørgensen FS, Oprea T, Brunak S, Taboureaux O (2008) hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol Pharmaceutics* 5:117
79. Li X, Zhang Y, Li H, Zhao Y (2017) Modeling of the hERG K⁺ channel blockage using online chemical database and modeling environment (OCHEM). *Mol Informat* 36:1700074
80. Thai K-M, Ecker GF (2008) A binary QSAR model for classification of hERG potassium channel blockers. *Bioorg Med Chem* 16:4107
81. Thai K-M, Ecker GF (2008) Classification models for hERG inhibitors by counter-propagation neural networks. *Chem Biol Drug Des* 72:279

82. Kireeva N, Kuznetsov SL, Bykov AA, Tsivadze AY (2013) Towards in silico identification of the human ether-a-go-go-related gene channel blockers: discriminative vs. generative classification models. *SAR QSAR Environ Res* 24:103
83. Liu L-l, Lu J, Lu Y, Zheng M-y, Luo X-m, Zhu W-l, Jiang H-l, Chen K-x (2014) Novel Bayesian classification models for predicting compounds blocking hERG potassium channels. *Acta Pharmacol Sin* 35:1093
84. Lu J, Lu D, Fu Z, Zheng M, Luo X (2018) Machine learning-based modeling of drug toxicity. In: Huang T (ed) *Computational systems biology: methods and protocols*. Springer, New York, p 247
85. Chavan S, Abdelaziz A, Wiklander JG, Nicholls IA (2016) A k-nearest neighbor classification of hERG K⁺ channel blockers. *J Comput Aided Mol Des* 30:229
86. Wang S, Sun H, Liu H, Li D, Li Y, Hou T (2016) ADMET evaluation in drug discovery. 16. Predicting hERG blockers by combining multiple pharmacophores and machine learning approaches. *Mol Pharmaceutics* 13:2855
87. Sun H, Huang R, Xia M, Shahane S, Southall N, Wang Y (2017) Prediction of hERG liability – Using SVM classification, bootstrapping and jackknifing. *Mol Inform* 36:1600126
88. Siramshetty VB, Chen Q, Devarakonda P, Preissner R (2018) The Catch-22 of predicting hERG blockade using publicly accessible bioactivity data. *J Chem Inf Model* 58:1224
89. Wacker S, Noskov SY (2018) Performance of machine learning algorithms for qualitative and quantitative prediction drug blockade of hERG1 channel. *Comput Toxicol* 6:55
90. Cavalli A, Poluzzi E, De Ponti F, Recanatini M (2002) Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of hERG K⁺ channel blockers. *J Med Chem* 45:3844
91. Durdagi S, Duff HJ, Noskov SY (2011) Combined receptor and ligand-based approach to the universal pharmacophore model development for studies of drug blockade to the hERG1 pore domain. *J Chem Inf Model* 51:463
92. Aronov AM, Goldman BB (2004) A model for identifying hERG K⁺ channel blockers. *Bioorg Med Chem* 12:2307
93. Aronov AM (2006) Common pharmacophores for uncharged human ether-a-go-go-related gene (hERG) blockers. *J Med Chem* 49:6917
94. Johnson SR, Yue H, Conder ML, Shi H, Dowejko AM, Lloyd J, Levesque P (2007) Estimation of hERG inhibition of drug candidates using multivariate property and pharmacophore SAR. *Bioorg Med Chem* 15:6182
95. Yamakawa Y, Furutani K, Inanobe A, Ohno Y, Kurachi Y (2012) Pharmacophore modeling for hERG channel facilitation. *Biochem Biophys Res Commun* 418:161
96. Kratz JM, Schuster D, Edtbauer M, Saxena P, Mair CE, Kirchebner J, Matuszczak B, Baburin I, Hering S, Rollinger JM (2014) Experimentally validated hERG pharmacophore models as cardiotoxicity prediction tools. *J Chem Inf Model* 54:2887
97. Durdagi S, Erol I, Salmas RE, Patterson M, Noskov SY (2017) First universal pharmacophore model for hERG1 K⁺ channel activators: actHER. *J Mol Graphics Model* 74:153
98. Cole JC, Korb O, Olsson TSG, Liebeschuetz J (2011) The basis for target-based virtual screening: protein structures. In: Sotriffer C (ed) *Virtual screening. Methods and principles in medicinal chemistry*. Wiley-VCH, Weinheim, p 87
99. Research Collaboratory for Structural Bioinformatics (2019) ProteinDataBank. <https://www.rcsb.org/>
100. Tribet C, Audebert R, Popot J-L (1996) Amphipols: polymers that keep membrane proteins soluble in aqueous solutions. *Proc Natl Acad Sci U S A* 93:15047
101. Shimada I, Ueda T, Kofuku Y, Eddy MT, Wüthrich K (2018) GPCR drug discovery: integrating solution NMR data with crystal and cryo-EM structures. *Nature Rev Drug Discov* 18:59
102. Renaud J-P, Chari A, Ciferri C, Liu W-T, Rémy H-W, Stark H, Wiesmann C (2018) Cryo-EM in drug discovery: achievements, limitations and prospects. *Nature Rev Drug Discov* 17:471

103. Venien-Bryan C, Li Z, Vuillard L, Boutin JA (2017) Cryo-electron microscopy and X-ray crystallography: complementary approaches to structural biology and drug discovery. *Acta Crystallogr* 73:174
104. Cavasotto CN, Phatak SS (2009) Homology modeling in drug discovery: current trends and applications. *Drug Discov Today* 14:676
105. Phatak SS, Gatica EA, Cavasotto CN (2010) Ligand-steered modeling and docking: a benchmarking study in class A G-protein-coupled receptors. *J Chem Inf Model* 50:2119
106. Blake JD, Cohen FE (2001) Pairwise sequence alignment below the twilight zone. *J Mol Biol* 307:721
107. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403
108. Schmidtke P, Ciantar M, Theret I, Ducrot P (2014) Dynamics of hERG closure allow novel insights into hERG blocking by small molecules. *J Chem Inf Model* 54:2320
109. Farid R, Day T, Friesner RA, Pearlstein RA (2006) New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies. *Bioorg Med Chem* 14:3160
110. Masetti M, Cavalli A, Recanatini M (2008) Modeling the hERG potassium channel in a phospholipid bilayer: molecular dynamics and drug docking studies. *J Comput Chem* 29:795
111. Stary A, Wacker SJ, Boukharta L, Zachariae U, Karimi-Nejad Y, Åqvist J, Vriend G, de Groot BL (2010) Toward a consensus model of the hERG potassium channel. *ChemMedChem* 5:455
112. Șterbuleac D, Maniu CL (2018) Computer simulations reveal a novel blocking mode of the hERG ion channel by the antiarrhythmic agent clofilium. *Mol Informat* 37:1700142
113. Aksoydan B, Kantarcioglu I, Erol I, Salmas RE, Durdagi S (2018) Structure-based design of hERG-neutral antihypertensive oxazolone and imidazolone derivatives. *J Mol Graph Mod* 79:103
114. Munawar S, Windley MJ, Tse EG, Todd MH, Hill AP, Vandenberg JI, Jabeen I (2018) Experimentally validated pharmacoinformatics approach to predict hERG inhibition potential of new chemical entities. *Front Pharmacol* 9:1035
115. Sottriffer C (2011) Virtual screening: principles, challenges, and practical guidelines, vol 48. *Methods and principles in medicinal chemistry*. Wiley-VCH, Weinheim
116. Sottriffer CA (2017) Protein-ligand docking: from basic principles to advanced applications. In: Cavasotto CN (ed) *In silico drug discovery and design: theory, methods, challenges, and applications*. CRC Press, Boca Raton, FL, p 558
117. Meng X-Y, Zhang H-X, Mezei M, Cui M (2012) Molecular docking: a powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des* 7:146
118. Sliwoski G, Kothiwale S, Meiler J, Lowe EW, Barker EL (2014) Computational methods in drug discovery. *Pharmacol Rev* 66:334
119. Li X, Li Y, Cheng T, Liu Z, Wang R (2010) Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes. *J Comput Chem* 31:2109
120. Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, Tian S, Hou T (2016) Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys Chem Chem Phys* 18:12964
121. Pagadala NS, Syed K, Tuszynski J (2017) Software for molecular docking: a review. *Biophys Rev* 9:91
122. Kitchen DB, Decomez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* 3:935
123. Liu J, Wang R (2015) Classification of current scoring functions. *J Chem Inf Model* 55:475
124. Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H-O (2005) Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J Med Chem* 48:2534
125. Jacobsson M, Lidén P, Stjenschantz E, Boström H, Norinder U (2003) Improving structure-based virtual screening by multivariate analysis of scoring data. *J Med Chem* 46:5781

126. Gao H, Williams C, Labute P, Bajorath J (1999) Binary quantitative structure–activity relationship (QSAR) analysis of estrogen receptor ligands. *J Chem Inf Model* 39:164
127. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 39:561
128. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 25:197
129. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijter MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KLH, Edwards DD, Shoichet BK, Roth BL (2009) Predicting new molecular targets for known drugs. *Nature* 462:175
130. Hecker N, Ahmed J, von Eichborn J, Dunkel M, Macha K, Eckert A, Gilson MK, Bourne PE, Preissner R (2012) SuperTarget goes quantitative: update on drug–target interactions. *Nucleic Acids Res* 40:D1113
131. Dunkel M, Günther S, Ahmed J, Wittig B, Preissner R (2008) SuperPred: drug classification and target prediction. *Nucleic Acids Res* 36:W55
132. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res* 35:D198
133. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiss A, Jensen LJ, Schneider R, Skoblo R, Russell RB, Bourne PE, Bork P, Preissner R (2008) SuperTarget and Matador: resources for exploring drug–target relationships. *Nucleic Acids Res* 36:D919
134. Gfeller D, Michielin O, Zoete V (2013) Shaping the interaction landscape of bioactive molecules. *Bioinformatics* 29:3073
135. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V (2014) SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res* 42:W32
136. O’Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. *J Cheminform* 3:33
137. Liu X, Vogt I, Haque T, Campillos M (2013) HitPick: a web server for hit identification and target prediction of chemical screenings. *Bioinformatics* 29:1910
138. Hamad S, Adornetto G, Naveja JJ, Chavan Ravindranath A, Raffler J, Campillos M (2018) HitPickV2: a web server to predict targets of chemical compounds. *Bioinformatics*:bty759-bty759
139. Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P (2012) STITCH 3: zooming in on protein–chemical interactions. *Nucleic Acids Res* 40:D876
140. Cheng F, Li W, Zhou Y, Shen J, Wu Z, Liu G, Lee PW, Tang Y (2012) admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J Chem Inf Model* 52:3099
141. Yang H, Lou C, Sun L, Li J, Cai Y, Wang Z, Li W, Liu G, Tang Y (2018) admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics*:bty707-bty707
142. Filimonov D, Poroikov V (2008) Probabilistic approaches in activity prediction. In: Varnek A, Tropsha A (eds) *Chemoinformatics approaches to virtual screening*. Royal Society of Chemistry, Cambridge, UK, p 182
143. Filimonov DA, Laguina AA, Glorizova TA, Rudik AV, Druzhilovskii DS, Pogodin PV, Poroikov VV (2014) Prediction of the biological activity spectra of organic compounds using the PASS online web resource. *Chem Heterocycl Compd* 50:483
144. Braga RC, Alves VM, Silva MFB, Muratov E, Fourches D, Tropsha A, Andrade C (2014) Tuning hERG out: antitarget QSAR models for drug development. *Curr Top Med Chem* 14:1399
145. Braga RC, Alves VM, Silva MFB, Muratov E, Fourches D, Lião LM, Tropsha A, Andrade CH (2015) Pred-hERG: a novel web-accessible computational tool for predicting cardiac toxicity. *Mol Inf* 34:698

146. Rossato G, Ernst B, Smiesko M, Spreafico M, Vedani A (2010) Probing small-molecule binding to cytochrome P450 2D6 and 2C9: an in silico protocol for generating toxicity alerts. *ChemMedChem* 5:2088
147. Vedani A, Dobler M, Smieško M (2012) VirtualToxLab – a platform for estimating the toxic potential of drugs, chemicals and natural products. *Toxicol Appl Pharmacol* 261:142
148. Vedani A, Dobler M, Hu Z, Smieško M (2015) OpenVirtualToxLab – a platform for generating and exchanging in silico toxicity data. *Toxicol Lett* 232:519
149. Kratz JM, Grienke U, Scheel O, Mann SA, Rollinger JM (2017) Natural products modulating the hERG channel: heartaches and hope. *Nat Prod Rep* 34:957
150. Wolber G, Langer T (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model* 45:160



Fabian Mayr studied pharmaceutical sciences at the University of Innsbruck, Austria, and graduated in 2017 with a Master's degree. He is currently working as a Ph.D. student at the Institute of Pharmacy, Department of Pharmacognosy and the Department of Pharmaceutical Chemistry at the University of Innsbruck. He is a member of the groups of Prof. Stuppner and Prof. Schuster. His main research interests are the search for lead structures from natural sources, as well as lead optimization of the latter—using both *in silico* and phytochemical techniques.



Christian Vieider studied pharmaceutical sciences at the University of Innsbruck, Austria, and obtained his Master's degree in 2018. His research interests are the *in silico* activity prediction and molecular modeling of the hERG channel. Currently he works as a pharmacist at the Paracelsus Pharmacy in Vipiteno, Italy.



Veronika Temml studied chemistry at the University of Innsbruck, Austria, and conducted work for her diploma thesis at the Max-Planck Institute for Biophysical Chemistry in Göttingen, Germany. After graduating in 2010, she returned to Innsbruck to undertake a Ph.D. in pharmaceutical sciences in the Computer-Aided Molecular Design Group of Prof. Schuster. She received her degree in 2014 and is currently a postdoctoral fellow at the Department of Pharmacognosy at the University of Innsbruck working on a Hertha Firnberg Fellowship, focusing on the discovery of multi-target anti-inflammatory compounds from Nature with *in silico* methods. Her research interests include lead optimization and activity and ADMET prediction with computational methods.



Hermann Stuppner is a trained pharmacist, and received his Ph.D. degree in pharmaceutical biology from the University of Munich. After a postdoctoral stay at the Department of Developmental and Cell Biology, Phytochemical and Toxicological Laboratory, University of California, he returned to the University of Innsbruck, Austria, for his habilitation in pharmacognosy. Since 2001, he has been a full professor at the Institute of Pharmacy/Pharmacognosy and, since 2003, the dean of Studies of the Faculty of Chemistry and Pharmacy. His main research interests are the isolation and structural elucidation of metabolites from higher plants, as well as their anti-inflammatory, antitumor, and antimicrobial activities, the analysis and quality assessment of medicinal plants and phytopharmaceuticals, the discovery of pharmaceutically active natural products using *in silico* techniques, and NMR-based metabolic profiling.



Daniela Schuster studied pharmacy at the University of Innsbruck, where she also received her Ph.D. degree in pharmaceutical chemistry. During her studies, she specialized in *in silico* bioactivity predictions of small organic molecules. Supported by an Erika Cremer habilitation fellowship, she completed her habilitation and later obtained one of the first Ingeborg Hochmair Professorships at the University of Innsbruck. Since 2018, she has been head of the Department of Pharmaceutical and Medicinal Chemistry at the Paracelsus Medical University in Salzburg, Austria. Her main interests are in teaching students the beauty of *in silico* medicinal chemistry and applying her knowledge and skills to drug discovery and environmental chemicals research.

A Strength-Weaknesses-Opportunities-Threats (SWOT) Analysis of Cheminformatics in Natural Product Research



Benjamin Kirchweger and Judith M. Rollinger

Contents

1	Introduction	239
2	S: Strengths of Cheminformatics in Natural Product Research	241
2.1	Availability and Access to Data	242
2.2	Natural Product Collections	244
2.3	Applicability of Cheminformatics Tools	245
3	W: Weaknesses of Cheminformatics in Natural Product Research	250
3.1	Structural Complexity of Natural Products	251
3.2	Handling of Glycosides	252
3.3	Tiny Databases	253
4	O: Opportunities of Cheminformatics in Natural Product Research	254
4.1	Virtual Screening of Natural Product Databases	255
4.2	Exploitation of Pharmacognostic Knowledge	256
4.3	Virtual Target Fishing	258
4.4	Binding Pose and Activity Predictions	259
5	T: Threats of Cheminformatics in Natural Product Research	259
6	Conclusion	260
	References	261

1 Introduction

Small molecule natural products are biosynthesized by biological systems to enable communication and interaction between cells, individuals, and species, serving as repellents, poisons, attractants, and signaling molecules. Owing to their biosynthetic enzyme origin and specific biological purposes, their chemical structures were designed by evolution to interact with macromolecules such as proteins, lipids, and nucleic acids [1–4]. This is in accordance with the finding of increased hit

B. Kirchweger · J. M. Rollinger (✉)

Department of Pharmacognosy, University of Vienna, Vienna, Austria
e-mail: benjamin.kirchweger@univie.ac.at; judith.rollinger@univie.ac.at

© Springer Nature Switzerland AG 2019

A. D. Kinghorn, H. Falk, S. Gibbons, J. Kobayashi, Y. Asakawa, J.-K. Liu (eds.),
Progress in the Chemistry of Organic Natural Products, Vol. 110,
https://doi.org/10.1007/978-3-030-14632-0_7

239

rates of natural product collections compared to synthetic and combinatorial collections in high-throughput screening campaigns [5, 6]. Analysis of such natural product collections revealed an exceptionally high diversity of molecular structures and properties, such as considerable molecular shape, stereogenic and ring-system complexity. They cover a broad chemical space, especially biologically relevant space [7–13] as outlined in detailed chapter “Cheminformatics Explorations of Natural Products” by Medina-Franco et al. in this volume (p. 1).

This makes natural products ideal candidates for drug discovery. Indeed, plants, fungi, and animals were almost the only source for pharmaceutical preparations for a long period of human history. Even with the advent of modern single-molecule medicines, natural products continued to play an important role [14]. A comprehensive analysis by Newman and Cragg points out that still 32% of all small-molecule approved drugs launched between 1981 and 2014 are unaltered natural products or natural product derivatives. Another 32% were inspired by natural products or their pharmacophores [15]. It is therefore tempting to speculate that natural product structures are privileged, possessing particular geometries; for instance, they exhibit a variety of novel, non-flat ring systems suitable for specific side chain substitutions, which are then prone to interact with an array of target proteins [16].

In contrast to the well-recognized high potential of natural products in drug discovery, the research engagement in this field has been dramatically scaled down in major pharmaceutical companies, mainly because it is stigmatized as an expensive endeavor [17]. The process of choosing a suitable biological source, its often limited or restricted access, the successful isolation of single active constituents from complex matrices, and deciphering their molecular structures seem too cumbersome compared to an increasingly automated and straightforward drug discovery process. New technologies like combinatorial chemistry, high-throughput screening using miniaturized and automatized assay batteries, and big data evaluation have triggered a great transformation in drug discovery [18]. The identification of ligands against specific targets as starting points for lead development is of utmost importance in this scientific field [19–21].

A major challenge in drug discovery from natural sources is hereby the identification of single bioactive constituents in order to establish unambiguous cause-effect relationships for later lead development. The classical approach for this task has been the bioassay-guided fractionation of crude or partly purified extracts [22]. Thus, a multicomponent mixture (extract) is separated step by step with subsequent assessment of the biological activities of the fractions obtained, followed by iterative rounds of separation and assaying [23–25]. Ideally, the goal is to end up with a single or a few purified active constituents—a goal which is certainly not often achieved because of certain shortcomings, such as insufficient robustness of bioassays used, potential solute adsorption to the solid phase during chromatographic fractionation, the re-isolation of previously known bioactive compounds, the failure to detect synergistic activity between the components present, and/or the decomposition of the constituents [14, 26].

The reverse path of testing pure natural products after isolation brings up several questions: (1) How to choose the natural starting organism? (2) Which components should be isolated? (3) How to choose a promising target for testing? Some of the most interesting natural products are difficult to isolate and only contained in small quantities in their natural source, for example, the yield of paclitaxel isolated from its source plant, *Taxus brevifolia* bark, was in the range of 0.01% [27]. Moreover, only 10% of all known natural products can be obtained by commercial suppliers [28], and these sometimes command very high prices. This is one reason their macromolecular targets remain largely unknown [29]. Natural products can be considered as too precious for dissecting their potential bioactivities by trial and error, and a rationale to streamline their biological evaluation is needed.

In this context, the application of *in silico* tools, in particular, virtual screening, has developed as an important strategy in natural product research for the prediction of ligand-target interactions and for rationalizing their bioactivity or even efficacy on a molecular level. Computational models can be created based upon already available information for the system under investigation and used to make predictions on new events. Without question, cheminformatics-based techniques are nowadays increasingly vital and substantial parts of modern-day drug discovery in medicinal chemistry, in both industry and academia. Their impact in natural product research is also increasing and has been reviewed elsewhere [30–32].

Here, we provide a comprehensive analysis of the strengths, weaknesses, opportunities, and threats (SWOT) of cheminformatics tools in natural product research. The analysis will provide a guide to facilitate their concatenation on the basis of past research projects, and aims to indicate gaps and caveats that exist. Therefore, the outcome of this analysis should give insight into strategic steps for further advances toward the combined use of cheminformatics and natural products drug discovery, to cope expediently with the challenges and opportunities in these two promising and prolific research areas.

2 S: Strengths of Cheminformatics in Natural Product Research

Cheminformatics is the use of computational and informational tools to understand and solve problems in the field of chemistry, particularly drug lead identification and optimization. The intended goal is to make better decisions faster [33]. In particular, virtual screening, which is the use of computational algorithms and models for the identification of bioactivities, has huge potential for more extensive application in natural product research [34, 35].

The implementation of cheminformatic tools can circumvent some of the costly and time-consuming bottlenecks prohibitive to drug discovery from natural sources.

From a pharmacognostic perspective, the prediction of molecular properties, possible targets but also antitargets of secondary metabolites, may be extremely useful to streamline experimental efforts, and hence to accelerate research and development projects. The scarce availability of isolated test materials demands for *in silico* predictions to unravel natural product molecular modes of actions and to deploy a rationale in lead development [32, 36–38]. From a cheminformatics perspective, virtual screening of collections consisting of fewer, but more sophisticated chemical entities, which are designed by evolution to interact specifically with macromolecular targets, rather than large synthetic molecule collections, can be a straightforward and prolific approach for the identification of novel lead compounds. The exploitation of natural product chemistry to implement Nature's privileged structures and chemical traits into synthetic compound repositories is another important topic [39–41].

From a retrospective analysis of research in the past two decades, the concatenation of cheminformatics and natural product research has certain prerequisites, which have gained substantial input and development in recent years, to categorize them as strengths. These refer to:

1. The availability and access to data providing available information on and the ability to obtain reliable data of the system under investigation
2. Natural product collections including their annotation to meta-data, curation, and a well-analyzed content
3. Availability and applicability of cheminformatic tools for the handling of natural products and specialized software and methods for event prediction

The following sections will provide more insight into the tools and most important databases available or literature dealing with this topic, without intending to provide a complete account.

2.1 Availability and Access to Data

A computational model's predictive power can be correlated roughly to the state of knowledge for the system it describes. The access to resources such as chemical databases, bioactivity collections, and biological data and a viable linkage and curation of these data is required to perform successful projects [42–44]. Lots of these resources are deposited and freely accessible. Chemical molecular databases with close to a billion virtual molecular entities have been established [44]. In 2017, four big chemical databases, PubChem, ChemSpider, Scifinder, and UniChem, compiled 95, 63, 134, and 154 million chemical structure records, respectively [45].

Biological and biomedical data stored in publicly available bioactivity databases provide a huge amount of detailed information on chemical entities in combination with target proteins, quantitative binding, and bioactivity values. The ChEMBL

database [46–48] connects 1.8 million 2D drug-like small molecule structure records with 12,000 molecular targets and 15.2 million bioactivities in an easily accessible interface. The data are derived mainly from seven medicinal chemistry journals (Bioorganic and Medicinal Chemistry Letters, Journal of Medicinal Chemistry, Bioorganic and Medicinal Chemistry, Journal of Natural Products, European Journal of Medicinal Chemistry, ACS Medicinal Chemistry Letters, MedChemComm) and selected articles from 200 journals and certain patents [48]. PubChem has compiled 239.6 million bioactivities for 3.4 million molecules, mainly from high-throughput screening experiments [49, 50]. Chemical patents represent another rich resource of chemical and biomedical information. The SureCHEMBL database aims to make the chemistry annotations of US, EP, WO, and JP patents available in a searchable interface. However, the connected biomedical data are not annotated [51, 52]. A smaller but highly curated database is the DrugBank with 12,000 chemical entries focusing on drugs and related molecules like nutraceuticals. Drug targets, pathways, indications and other pharmacological information are provided [53–55]. A large and comprehensive biomedical database of natural products does not yet exist. The Protein Data Bank (PDB) is a valuable resource for 3D information on biological macromolecules. It archives 144,000 experimentally determined structures and their complexes with metals, co-factors, crystal water, and small-molecule ligands [56, 57].

Table 1 summarizes the most important free accessible databases of biomedical and biological information useful in cheminformatics. A more detailed list has been compiled in [58]. It should be noted that the quality of information in the databases differs due to diverse data sources, data acquisition procedures, and curation efforts.

Chemical, biomedical, and other life science data can be estimated to grow further in the future as the integration of chemical information from multiple sources and analytical techniques, extracting and mining information from journal articles and patents is still improving. Collaborative efforts and the commitment to make generated data available in the public domain will stimulate this development.

Table 1 Biomedical databases

Database	Content	Size	References
BindingDB	Experimental protein-small molecule binding affinities	1.2 million binding data for 55,000 proteins and 520,000 drug-like molecules	[59]
CHEMBL	Data compiled from literature; PubChem and SureChEMBL	1.8 million drug-like small molecules 15.2 million bioactivities	[46–48]
Drugbank	Highly curated drug data combined with drug target, pathway, indication, and other pharmacological information	12,000 nutraceuticals, approved and experimental drugs	[53–55]
DUD.E.	Active compounds and target affinities, includes widely used decoys in virtual screening	22,886 actives 102 targets 50 decoys for each active	[60, 61]
GLASS	Manually curated repository for experimentally validated GPCR-ligand interactions	342.5 million ligands 3 million GPCR targets	[62]
GOSTAR	Manually curated SAR database	6.6 million inhibitors 22 million quantitative SAR points	[63]
OCHEM	ADME data	2.8 million property records	[64, 65]
PDBbind	Binding affinities of PDB entries	11,000 binding affinities	[60]
PubChem	Chemical database with bioactivity data from HTS assays	63 million molecules For 3.4 million molecules 239.6 million bioactivities are compiled	[49, 50]
Binding MOAD	High-quality PDB subset of ligand-protein complexes	33,000 structures	[67, 68]
PDB	Databank of experimentally determined structures of proteins, nucleic acids and complex assemblies	144,000 experimental determined macromolecule structures	[56, 57]
SMPDB	Interactive and visual small molecule pathway database	30,000 human pathways	[69, 70]
TTD	Database of therapeutic targets	3000 targets	[71]

DUD.E database of useful decoys, *GLASS* GPCR-ligand association database, *GOSTAR* global online structure-activity relationship database, *GPCR* G-protein-coupled receptor, *MOAD* mother of all databases, *OCHEM* online chemical database, *PDB* protein databank, *SAR* structure-activity relationship, *SMPDB* small molecule pathway database, *TTD* therapeutic target database

2.2 Natural Product Collections

A prerequisite of conducting cheminformatics in natural product research is the existence of stereochemically well-defined molecules. Appropriate commercial and also free natural product databases are available. These important resources have been reviewed several times [28, 72–77].

The most comprehensive database is the *Dictionary of Natural Products* (DNP) with currently 260,000 natural products. Information on trivial names, physicochemical properties, and toxicity data are supplied. For pharmaceutical biologists the information on biological sources and experimental properties such as UV spectra and dissociation constants can be very useful. Caution should be given when used for 3D applications, because the stereochemistry is not annotated in the 2D connection tables. The database was built manually by a team of academics and freelancers, who enable reconciling of errors and ensure high quality data [28, 78]. Although this database is comprehensive and well curated and covers a large chemical diversity, its availability only on a commercial basis hampers its broader use by the interested scientific community.

Alternatives are free virtual natural product collections, like the Universal Natural Product Database (UNPD), the TCM database@Taiwan, NPCARE, and the NuBBE database; these all have been made available free of charge [79–83]. Chen et al. recently have analyzed the content of natural product collections and observed a large overlap (108,000 molecules) of free virtual natural product collections with the DNP [28]. A thorough survey on natural product resources and their characteristics is provided in the chapter “Resources for Chemical, Biological, and Structural Data on Natural Products” by Kirchmair et al. in this volume (p. 37).

The use of cheminformatics tools to select natural products and natural product like compounds from large chemical (e.g., PubChem [50]), biomedical (e.g., ChEMBL [47], PDB [57]) or commercial vendor databases (e.g., ZINC [84], Aldrich Market Select [85]) would be a worthwhile strategy. Several tools able to identify natural products in large molecule sets have been developed. They are based on different machine learning tools such as rule-based approaches, similarity measurements of structural space, or connectivity fingerprints [86–91]. Recently, a random forest classifier with high accuracy was made available in a free online tool [92].

The diligent exploitation of natural product resources from widely unexplored organisms from different niches of our globe and the closer examination of already investigated marine and terrestrial organisms by advanced technical means will continue to extend the diversity and coverage of natural product collections. The exchange of virtual physically available collections between cooperation partners has been suggested to increase the access to natural products [93]. Efforts to compile, annotate, analyze, and finally enable their availability to a broad community will lead to an increasingly valuable resource for future drug discovery.

2.3 *Applicability of Cheminformatics Tools*

As summarized earlier, scientists have to learn from the vast amount of biomedical data generated and made available via data-sharing platforms. However, it is indisputable that the amount of data is far beyond traditional analysis and learning [94]. To create predictive cheminformatic models from big data, various approaches have been established ranging from comprehensive similarity measurements (e.g.,

pharmacophore, shape-based approaches, physicochemical property comparison) to complex molecular docking and sophisticated machine learning approaches (e.g., self-organizing maps). The basic concepts underlying these methods have been reviewed elsewhere [30–32, 95].

Notably, most virtual screening, binding pose prediction, and target fishing approaches have been shown to be also applicable to natural products. From the examples presented in Tables 2, 3, and 4, previous studies have been carried out frequently with user-friendly comprehensible *in silico* tools. Three-dimensional pharmacophore alignments, e.g., with Catalyst or LigandScout, and molecular docking, e.g., with Autodock Vina or Glide, offer an intuitive interface and allow easy implementation also to scientists not specialized in cheminformatics. These

Table 2 Approaches for the prediction of natural product binding modes

Technique	Software	Target ^a	Target class ^b	Examples
Molecular docking	Ligandfit	HR	V	[96]
	GOLD	NA	V	[97]
		5-LOX	E	[98]
		COX-2	E	[98]
		11 β -HSD1	E	[99]
	Glide	MD-2	PPI	[100]
		5-HT _{2C}	GPCR	[101]
	MOE	PPAR γ	TF	[102]
	Autodock	AChE	E	[103]
	Autodock Vina	NF- κ B	TF	[98]
CDOCKER	PPAR γ	TF	[104]	
Molecular dynamic simulation	AMBER	NA	V	[97]
		MD-2	AG	[100]
		DNA	DNA	[105]
		AChE	E	[103]
	NAMD	NF- κ B	TF	[98]

^aTarget abbreviations: *11 β -HSD1* 11 β -hydroxysteroid dehydrogenase type 1, *5-LOX* 5-lipoxygenase, *5-HT_{2C}* 5-hydroxytryptamine_{2C} receptor, *AChE* acetylcholinesterase, *COX-2* cyclooxygenase-2, *DNA* deoxyribonucleic acid, *HR* human rhinovirus coat protein, *MD-2* lymphocyte antigen 96, *NA* neuraminidase, *NF- κ B* nuclear factor kappa-light-chain-enhancer of activated B cells, *PPAR γ* peroxisome proliferator-activated receptor gamma

^bTarget class abbreviations: *AG* antigen, *DNA* deoxyribonucleic acid, *E* enzyme, *GPCR* G-protein coupled receptor, *PPI* protein-protein interaction, *TF* transcription factor, *V* viral protein

Table 3 Different approaches for the prediction of natural product molecular targets

Technique	Strategy/software	Examples
Artificial neural networks	Self-organizing maps, e.g. [106]	[29, 36, 107]
Hierarchical clustering	Based on <i>in silico</i> retrobiosynthesis [108]	[109]
Virtual parallel screening	Ligandprofiler, PipelinePilot, Ligandscout, Catalyst	[95, 110, 111]
Reverse docking	Autodock Vina	[112, 113]

Table 4 Different/complementary virtual screening approaches applied to natural products

Approach	Strategy/software	Target ^a	Target class ^b	Examples
Pharmacophore-based virtual screening	Catalyst	AChE	E	[114]
		COX-1, COX-2	E	[115, 116]
		HR	V	[96]
		hERG	IC	[117, 118]
		FXR	TF	[119, 120]
		cPLA ₂ α	E	[121]
		mPGES-1	E	[122]
		IKK- β	E	[98]
		mGlu	GPCR	[123]
	PrPC	V	[124]	
	Ligandscout	AChE	E	[114]
		hERG	IC	[117, 118]
		GPBAR1	GPCR	[125]
		11 β -HSD1	E	[99, 126]
		PPAR γ	TF	[127]
		CETP	LTP	[128]
	PharmaGIST	AMA1-RON2	PPI	[129]
	MOE	PPAR γ	TF	[102]
TbGAPDH		E	[130]	
2D similarity search	chemGPS [7]	Antichlamydial	–	[131]
	Connectivity fingerprints	FXR	TF	[132]
3D similarity search	ROCS	GPBAR1	GPCR	[125]
		NA	V	[133]
		IKK- β	E	[134]
	SQUIRREL	mPGES-1	E	[135]
	Phase	HIV-1 RT	V	[136]
Molecular docking	Autodock	Complex III	E	[137, 138]
		NA	V	[139]
		AMPK	E	[140]
	Autodock Vina	ROCK1	E	[141]
		Complex III	E	[137, 138]
	GOLD	AChE	E	[142]
		CK2	E	[143]
	Glide	HIV-1 RT	V	[136]
		CK2	E	[143]
		FXR	TF	[132]
		PPAR γ	TF	[144]
		Sirt1	E	[145]
		ACE	E	[146]
	LigandFit	mGlu	GPCR	[123]
	CDOCKER	PrPC	V	[124]

(continued)

Table 4 (continued)

Approach	Strategy/software	Target ^a	Target class ^b	Examples
	MOE	CK2	E	[143]
		<i>TbGAPDH</i>	E	[130]
	Molsoft	TNF- α	PPI	[147]
		DNA	DNA	[148]
	DOCK	AMPK	E	[140]
ROCK1		E	[141]	
QSAR	GP regression ^c	IRF-7	TF	[149]
	Multiple linear regression	Antitrypanosomal	–	[150]
Machine learning	Self-organizing maps	AChE	E	[142]
	Random forest classifier	AMA1-RON2	PPI	[129]
	GP regression ^c	PPAR γ	TF	[151]

^aTarget abbreviations: *11 β -HSD1* 1 β -hydroxysteroid dehydrogenase type 1, *5-LOX* 5-lipoxygenase, *ACE* angiotensin-converting enzyme, *AChE* acetylcholinesterase, *AMA1* apical membrane antigen 1, *AMPK* 5' AMP-activated protein kinase, *CETP* cholesteryl ester transfer protein, *CK2* casein kinase 2, *Complex III* coenzyme Q-cytochrome c-oxidoreductase, *COX-1* cyclooxygenase-1, *COX-2* cyclooxygenase-2, *cPLA₂ α* Cytosolic phospholipase A2 α , *DNA* deoxyribonucleic acid, *FXR* farnesoid X receptor, *GPBAR1* G protein-coupled bile acid receptor, *hERG* human ether-à-go-go-related gene potassium ion channel, *HIV-1 RT* human immunodeficiency virus type 1 reverse transcriptase, *HR* human rhinovirus coat protein, *IKK- β* inhibitor of nuclear factor kappa-B kinase subunit beta, *IRF-7* Interferon regulatory factor 7, *mPGES-1* microsomal prostaglandin E synthase-1, *MD-2* lymphocyte antigen 96, *NA* neuraminidase, *NF- κ B* nuclear factor kappa-light-chain-enhancer of activated B cells, *mGlu* metabotropic glutamate receptor, *PPAR γ* peroxisome proliferator-activated receptor gamma, *PrPC* cellular prion protein, *ROCK1* Rho-associated protein kinase, *RON2* rhopty neck protein 2, *Sirt1* NAD-dependent deacetylase sirtuin-1, *TbGAPDH* *Mycobacterium tuberculosis* glyceraldehyde-3-phosphate dehydrogenase, *TNF- α* tumor necrosis factor ligand superfamily member 2

^bTarget class abbreviations: *DNA* deoxyribonucleic acid, *E* enzyme, *GPCR* G protein-coupled receptor, *IC* ion channel, *LTP* lipid transfer protein, *PPI* protein-protein interaction, *TF* transcription factor, *V* viral protein

^cGaussian process regression

methods are already well-established and have demonstrated solid performance as shown by many successful projects [98, 99, 101, 123, 126, 134, 137, 138]. Most studies have combined different methods such as molecular docking and molecular dynamic simulations for the prediction of binding modes [100] or shape and molecular docking for virtual screening [136]. Further cheminformatic approaches, such as artificial neural networks, increasingly have gained importance, especially for target and activity prediction [29, 36, 107], and also in qualitative virtual screening experiments [129, 142, 151] (see the chapters “The Pharmacophore Concept and Its Applications in Computer-Aided Drug Design” and “Cheminformatic Analysis for Natural Product Fragments” of Langer and Reker, this volume (p. 97 and 141)).

Perhaps the most important cheminformatics application for natural product researchers is the prediction of molecular targets as thoroughly reviewed in the chapter “A Toolbox for the Identification of Modes of Action of Natural Products” provided by Rodrigues et al. (this volume, p. 73). Besides virtual target fishing of new isolates, it can help to fast forward the rationalization of traditionally used herbal remedies, the prediction of side effects, and the profiling of polypharmacologic actions [29, 30, 110, 112, 152]. The experimental validation of the target-predicting approaches is usually demonstrated on single molecules or only on few examples rather than on a large set of natural products [36, 108, 113], mainly owing to the major effort necessary for experimental testing and the limited physical availability of compounds.

The benefit of experimental testing based on virtual predictions compared to serendipitous experimental screening could be demonstrated convincingly by Doman et al. [153]. Their random *in vitro* screening for protein tyrosine phosphatase inhibitors revealed a hit rate of 0.02%, while assaying the virtually predicted hits yielded a hit rate of 34.8%. In general, the first evaluation of virtual hits does not require any physically available material but requires a critical check on various parameters before compounds are selected as candidates for experimental testing, e.g., availability; isolation efforts; physicochemical parameters referring to PAINS or inappropriate absorption, distribution, metabolism, excretion, and toxicity (ADMET); reported toxicity; and reliability of predictions [72, 30]. Rare biological material and precious isolates can be saved, and fewer bioassays are needed for the identification of active hits.

Computer-aided techniques have shown to be applicable to many natural product scaffolds such as polyketides [109], alkaloids [37, 118], coumarins [111, 125], flavonoids [133], and sesqui- and triterpenes, [99, 126, 150], and they have been used to make predictions on many biological drug target classes and phenotypic effects.

The concatenation of cheminformatics tools in combination with pharmacognostic expertise and complementary empirical knowledge, such as information from traditional medicine, *in vivo* studies, epidemiological or clinical investigations, bioassay-guided fractionation, and high-resolution mass spectrometry-based dereplication is able to dramatically enhance the true positive hit rates as discussed in Sect. 4.2 [116, 117].

The ever-increasing computing power and availability of augmented data analysis algorithms have led to a broad use of computational tools in drug discovery. Even big data quantities can be processed with increasingly clever algorithms. Moreover, some predictive methods have shown similar performance levels to a group of experienced medicinal chemists in predicting biological activities, and outperform the brains of experts in the ability to process large databases [154].

3 W: Weaknesses of Cheminformatics in Natural Product Research

The many successful projects documented in the literature should not lead to wrong perceptions. The processing of natural products with cheminformatics bears some caveats, risks and limitations, which are present not only in both domains (cheminformatics and natural product research) but also at their interface (Fig. 1). To overcome weaknesses, these limitations should be recognized in order to be considered and avoided as far as possible.

The limited availability of natural starting material [155] and of readily available natural products by commercial vendors [28], the absence of elucidated molecular structures for the vast majority of natural products that exist, in addition to assay interference [156], are examples of drawbacks with respect to natural products. The complexity of multicomponent mixtures with difficult-to-predict additive effects and separation problems in isolation efforts are further caveats.

In the field of cheminformatics, there are recommended reviews on the pitfalls of virtual screening [157] also in combination with natural product research [72, 30]. The most fatal weakness of cheminformatics approaches is that they have an inherent incapability to find novel compounds or novel molecular mechanisms of action. They can just extend knowledge on existing topics; the predictive power is better the more knowledge is available already for the system under investigation. An investigator has to navigate on the one hand between innovation usually

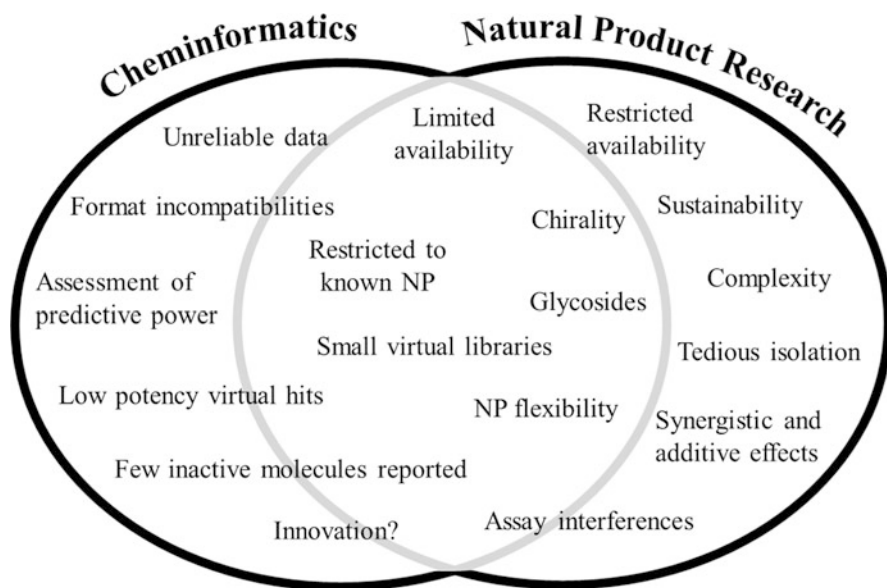


Fig. 1 Weaknesses and challenges in cheminformatics, in natural product research, and at the interface of these two fields

combined with interesting but ambitious topics with few relevant data available, and, on the other hand, probably trite, less risky targets with good prospects of success due to a vast amount of information already available. A number of molecular mechanisms have been explored by means of natural products, and some biological targets have even been named by their natural product ligands, as exemplified by the muscarinic acetylcholine receptor and cannabinoid receptors. Therefore, *in silico* tools should be used as part of an interconnected network combined with empirical knowledge and phenotype-directed and target-directed screening platforms [38, 158].

3.1 Structural Complexity of Natural Products

A main weakness appearing upon the handling of natural products with computational algorithms is the difference between natural products and synthetic small molecules [35], which was previously analyzed by several groups [11, 12, 76, 159–161]. Most algorithms were trained on synthetic molecules and might perform less well when they are confronted to unfamiliar molecules [35].

Natural products differ from other compound sets in several molecular properties. They are more hydrophobic and contain more oxygen atoms and fewer nitrogen atoms compared to synthetic drugs. The structural complexity, especially the differences in ring architecture with unsaturated ring systems and more three-dimensional molecular shapes but less aromaticity is, on the one hand, closely correlated to the concept of privileged structures but may cause a difference in performance [161].

Natural products are more flexible due to high numbers of sp^3 hybridized atoms making computations with three-dimensional tools (e.g., molecular docking) and conformational sampling for 3D similarity searches or pharmacophore-based virtual screening slower and more error-prone. A large number of rotatable bonds can also lead to promiscuous results, where ligands are fitted to molecular shapes, pharmacophores, and molecular docking in implausible ways. Rotatable bond filters for shape matching experiments like the suggested Veber rule (rotatable bonds <12) [162] can be applied.

A characteristic of natural products is the frequent occurrence of one or even more chiral centers [11, 76, 160], which are not always annotated in natural product databases or catalogs of chemical vendors [78]. Moreover, the exact configuration is not always reported in the primary literature. The generation of all stereochemical configurations is time-intensive and error-prone.

Projects are more likely to be successful if the input information is related to the test subjects. Screening of natural product collections with a synthetic molecule query may be problematic concerning the reliability of the prediction. Similarly, the screening of synthetic molecule collections with a natural product-like query may lead to disappointing results. It is obvious that different ligands can occupy different regions on the same protein, even in the same binding site, making 3D alignments

like pharmacophore- and shape-based screening prone to high rates of false-negative results [157].

3.2 *Handling of Glycosides*

Glycosides play an important role in living organisms and are abundant moieties of natural products with different biological roles. Glycosides like amygdalin are used by different plants as storage and transport forms of their aglycone molecules. Upon disruption of compartments (e.g., by grazing herbivores), enzyme hydrolysis cleaves the glycosides and sets free the toxic aglycone. Other glycosides are natural prodrugs, enabling improved drug-likeness of the transformed metabolites [163, 164].

At first glance and from a medicinal chemistry perspective, sugars and sugar-like moieties are not in the focus of drug discovery. They are easily cleaved in the gut by microbes or by first-pass metabolism, increase the molecular weight, and lead to steric hindrance. Further, the polar glycoside moiety hinders the lipophilic effect between protein and ligand. Therefore, algorithms were created to cleave sugars from their aglycone counterparts for creation of virtual screening databases [28, 165].

The molecular docking force field was adjusted to the binding of comparably rigid and nonpolar molecules and performs therefore well on such molecules; however, the performance with carbohydrates and carbohydrate-containing molecules is questionable. The frequently used molecular docking tool Autodock Vina was able to produce acceptable structures within the top five ranked poses in only 55% of experimental crystallographic carbohydrate-protein complexes [166].

Notably, glycosides have been important drugs for a long time. In herbal medicines, it is acknowledged that glycosides decrease capillary fragility and exert secretolytic, diuretic, and antiexudative effects [167–169]. Carbohydrates play important biological roles such as cell signaling, infection, and protein function [170–172]. These effects are mediated generally by nonclassical modes of action such as membrane activity and interaction with protein surfaces yet difficult to describe with algorithms [173–175].

There are also examples of classic ligand-target interactions with natural product glycosides. Thus, phlorizin, a dihydrochalcone derivative, was the blueprint for sodium-dependent glucose transporter 2 inhibitors. The sugar moiety of phlorizin represents a vital part of the necessary pharmacophore to block the transporter [176]. From perspectives such as this, it may be a fallacy to exclude glycosides from virtual screening databases.

The handling of glycosides may be dependent on the individual target and project but definitely needs consideration. Further improvement of virtual screening tools toward a better applicability for glycosides is certainly needed.

3.3 *Tiny Databases*

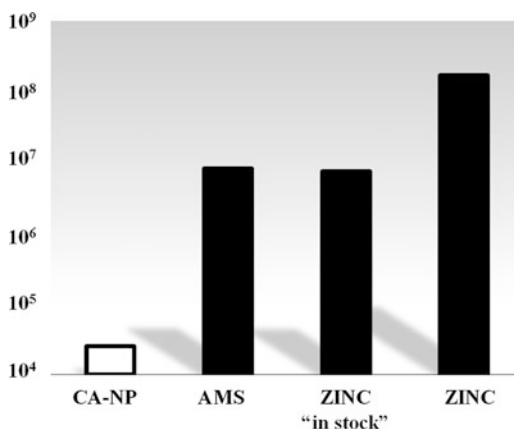
A comparison of commercially available natural product collections with synthetic collections reveals a large difference in their size (Fig. 2). When compared to large databases of commercially available synthetic and mixed collections like Aldrich Market Select [85] with 8 million unique available compounds, and ZINC [84] comprising 120 million available compounds (7.3 million in stock), 11,000 natural products available from natural product-only catalogues and 25,000 natural products in total (including natural products in mixed catalogues) are fairly small [28]. In total, an estimated 250,000–300,000 natural products are known up to now [28, 83].

Model rigidity has to be balanced according to the size of the databases screened. Assuming a restrictive model with a hit rate of 0.2% will lead to estimated 50 virtual hits from commercially available natural product databases and 16,000 virtual hits from commercially available synthetic molecules.

Natural product chemical diversity, however, is insufficiently explored and is biased toward molecules from extensively exploited sources making a final statement on their extent speculative. This is underlined, for example, by the discovery of naturally occurring organohalogens, which were considered until quite recently as rare and exotic isolates and often suspected to be artifacts. With the exploitation of unexplored sources such as marine organisms, algae, and lichens, thousands of these have been described [177]. Also, improved isolation and analytical methods, which enable the characterization of natural products contained at even lower traces, constantly change our perception of natural products chemistry.

Two main issues in future will be to continue the present rate of natural product discovery and to properly exploit what is found [178].

Fig. 2 Amounts of purchasable compounds in virtual collections on a logarithmic scale; white, natural product; black, primarily synthetic molecules; CA-NP, commercially available natural products; AMS Aldrich Market Select



4 O: Opportunities of Cheminformatics in Natural Product Research

The growing popularity in the usage of computer-aided techniques in natural product research resulted in numerous successful application examples. Depending on the scientific issues at hand and the available information, in addition to that missing, different in silico tools and strategies have to be carefully selected. Figure 3 provides a schematic overview on opportunities to approach scientific questions by cheminformatic means. Besides the individual application examples named in Tables 2, 3, and 4, some successful projects are outlined in this chapter.

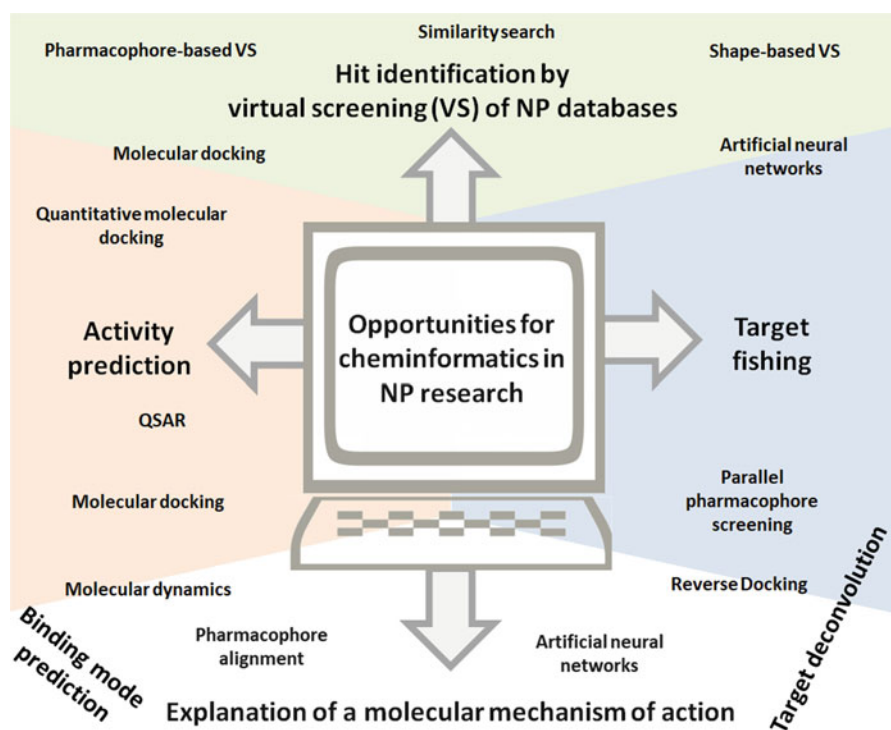


Fig. 3 Opportunities and areas of applications of cheminformatics in natural product research

4.1 Virtual Screening of Natural Product Databases

When considering the innate character of natural product collections (prolific, but low number of entities, difficult availability, high cost to obtain, etc.), as discussed in the previous chapters, it is highly recommended to first validate the predictive power of the model used by experimental testing of a set of virtual hits from easily accessible and inexpensive, physically available (synthetic) databases. Also, a proper preparation of the database subjected to virtual screening, e.g. by pre-connected filtering experiments may help to (1) focus on the most interesting candidates and (2) economize computational power.

For example, Su et al. prepared a virtual screening collection with fingerprint clustering and drug-likeness filters. Natural products unsuitable for the molecular docking algorithms due to their size and polarity could be removed in advance. The virtual screening of only 24,000 molecules with a stepwise workflow employing molecular docking led to the identification of baicalein and phloretin as new natural Rho kinase inhibitors [141].

Considerable database preparation was also performed by Costa et al. for the identification of HIV-1 reverse transcriptase inhibitors. They generated a natural product database from 11 vendors and natural product databases publicly available in the ZINC repository. They narrowed down the database by removing molecules violating the Lipinski Rule of Five [179], and with predicted poor solubility and permeability. A parallel molecular docking protocol as well as a 3D similarity search led to the selection and experimental testing of several virtual hits. β -Carboline derivatives were identified as HIV-1 reverse transcriptase inhibitors and their binding mode was examined using the molecular docking predictions as well as molecular dynamic simulations [136].

Insufficient capacities to obtain large sets of natural products for experimental testing may be circumvented by the application of a set of ligand-based pharmacophore models previously validated mainly on synthetic molecules for the most prevalent antitarget in drug discovery and development, i.e., the hERG channel [38, 117, 118]. For a detailed insight into the performance of different hERG prediction tools toward a fast and efficient cardiotoxic risk assessment, reference is made in the contribution in the chapter “Open Access Activity Prediction Tools for Natural Products. Case Study: hERG Blockers” by Schuster (this volume, p. 175). Kratz et al. used the previously generated, best performing pharmacophore model for the subsequent virtual hERG screening of natural product databases. They validated their predictions in a patch clamp assay by testing small-scale lead-like enhanced extracts from 12 plant materials known to contain the virtual hits. At 100 $\mu\text{g}/\text{cm}^3$, 4 out of the 12 extracts exerted a hERG tail current inhibition of more than 30%, among them *Ipecacuanhae Radix*. Use of an appropriate phytochemical workflow resulted in the isolation and identification of five out of the six virtually predicted alkaloids, among them the major constituents emetine and cephaeline with IC_{50} values of 21.4 and 5.3 μM , respectively [118]. Similarly, Vuorinen et al. [126] used

pharmacophore models for the identification of hydroxysteroid dehydrogenase inhibitors from Nature using previously validated models [180, 181].

Virtual screening can also predict phenotypic efficacy as shown by work of Karhu et al. [131]. They performed a principal component analysis [7] of the physicochemical properties of a natural product database and an antichlamydial reference set and compared the Euclidian distances in the chemical space. Out of 26 virtual hits, 6 molecules were confirmed as active and 1 high-potency lead was identified.

4.2 Exploitation of Pharmacognostic Knowledge

The implementation of information from traditional medicine and the knowledge from structural ligand-target interaction can increase significantly the yield of true active hits (Fig. 4a, b). Applying pharmacophore models for cyclooxygenase (COX) inhibitors, which were completely derived with input from synthesis chemistry, Rollinger et al. were able to demonstrate statistically their effectiveness in the field of natural products. A comparison of virtual hits obtained by screening of the mainly synthetic molecular 3D collections of the Derwent World Drug Index (WDI) and the Database of the National Cancer Institute (NCI) revealed hit rates in the range of 6.6% to 13.7% (depending on the search queries used). Using the in-house-generated natural product database NPD consisting of molecular structures from 80,000 natural products, even a slight increase of molecules that virtually fit into the

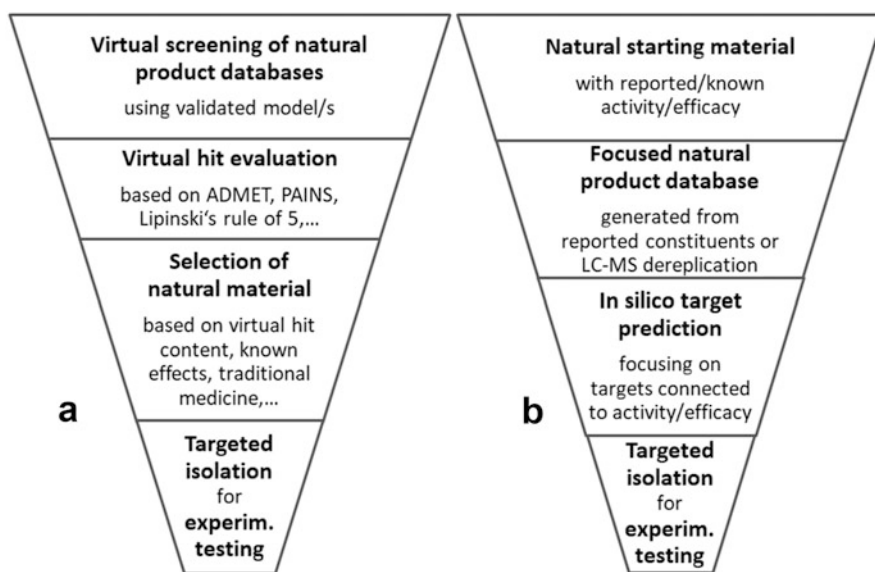


Fig. 4 Examples of strategies for the implementation of cheminformatics in pharmacognostic workflows: (a) starting from validated in silico model/s; (b) starting from bioactive natural material

required features of the pharmacophore models could be achieved. A striking result of this study, however, was the average increase of hit rates (77 to 133%) when an ethnopharmacologically biased database labeled as DIOS was screened compared to the WDI and the NCI. The DIOS database contains structural information of 28,000 secondary metabolites reported from those medicinal plants that Pedanius Dioscorides (first century AD) described in his “De Materia Medica” as useful in the application of different sorts of inflammation. In this way, the distinct statistical benefit of a combination of an ethnopharmacological approach and an *in silico* screening could be demonstrated [115, 116]. In a follow-up study, one of the most promising herbal drugs, the root bark of *Morus alba*, was selected based on the predictions from the DIOS database. The plant material was phytochemically investigated to evaluate the applicability of the computer-aided approach. Several virtually predicted constituents from the group of the isolated Diels-Alder adducts could be confirmed successfully as COX inhibitors [116].

Kirchweger et al. [125] performed a virtual screening of several small natural product databases and a larger synthetic small molecule collection (SPECS) for the identification of activators for the G protein-coupled bile acid receptor 1 (GPBAR1) using a ligand-based pharmacophore virtual screening approach. The virtual hits were ranked according to a shape-focused similarity score and the molecules were clustered according to their physicochemical properties. This approach enabled the selection of chemically diverse compounds endowed with the putative structural requirements to act as ligands of the envisaged target for experimental validation using a reporter-gene based assay. Both synthetic and natural product-derived virtual hits were subjected to experimental testing. Accordingly, the yield of active synthetic compounds (>15% receptor activation at 20 μM) was 10.5% (2 out of 19); natural products resulted in a five-time higher hit rate (57%; 8 out of 14). The latter group also included two novel GPBAR1 activating scaffolds, namely, the sesquiterpene coumarins farnesiferol B and microlobidene, which at 20 μM increased the receptor activation to 61% and 84%, respectively, thus showing an activity comparable to that of the endogenous ligand, lithocholic acid.

Cheminformatics can also be used in a straightforward manner for the identification of active principles of traditionally used medicines and unravel their molecular modes of actions. Schuster et al. [120] generated a set of validated pharmacophore models for the transcription factor FXR, a drug target for inflammatory liver diseases [182]. Grienke et al. [119] used this model for virtual screening of the Chinese herbal medicine database, and, from this work, lanostane-type triterpenes from the fruit body of *Ganoderma lucidum* were predicted as virtual hits. As this mushroom is traditionally used against hepatitis, liver disease, and arthritis, a full mycochemical investigation and isolation was performed. Five isolated lanostane triterpenes were confirmed experimentally to induce FXR activation with EC_{50} values in the low micromolar range.

4.3 *Virtual Target Fishing*

It is a frequent observation that a herbal drug shows a well-documented biological or clinical effect, but the constituents responsible as well as their underlying mechanisms of action remain elusive [95, 108]. Binding mode prediction and virtual target fishing can help to fast forward the rationalization of research and identify possible drug leads. Similar to already described nutritional and medicinal effects in humans, an observed phenotypic effect such as cytotoxicity, antimicrobial, or hypoglycemic activity can be followed up with focused isolation and experimental efforts.

In 2014, Reker et al. [29] presented a novel method for target fishing, which is independent of the target structure. The approach uses topological pharmacophore features of query compound fragments to compare them to pre-calculated drug compound clusters. The constituent can then be assigned to the cluster with the smallest Euclidian distance. Target information for the cluster was derived from confirmed interaction partners of reference drugs within the cluster. As a prospective application example, the macrolide archazolid A (ArcA) was investigated. This compound exerts potent cancer-related effects by inhibiting the ion pump vacuolar-type H⁺-ATPase at the nanomolar level. However, it was suggested that additional targets might be responsible for the pronounced antitumor effect. The analysis predicted several targets involved in arachidonic acid-associated signaling cascades as potential interaction partners, and subsequent biological testing confirmed a concentration-dependent effect of ArcA on half of these targets. In addition, weak effects on two further targets were observed. The experimental results validated the applicability of the natural product-derived fragment-based approach for the identification of novel macromolecular targets. Remarkably, all newly identified interaction partners of ArcA have also been linked to putative anticancer effects [29].

Mastic gum has been used traditionally against metabolic disorders [183] and has also shown to exert a hypoglycemic *in vivo* activity [184]. Its bioactive constituents and the molecular targets responsible were largely unknown. The virtual screening of a natural compound database against 11 β -HSD1 pharmacophore models retrieved triterpenoids from *Pistacia lentiscus* as virtual hits. Together with empirical and preclinical data, the prediction seemed plausible. Therefore, mastic gum and its acidic fraction, which is known to contain the predicted hits, were subjected to experimental testing. Both samples inhibited 11 β -HSD1 in a concentration-dependent manner; the two virtually predicted main triterpenes showed *IC*₅₀ values in the low micromolar range [126].

Gong et al. [112] used a similar approach based on reverse docking against 211 cancer-related targets to explain an observed cytotoxic effect against two cancer cell lines of two novel sponge metabolites. The precious isolates were only tested against the two most promising targets according to the docking scores. The experimental testing explained the phenotypic effects as attributed to the inhibition of histone acetyltransferase h(p300).

Several target prediction tools have been made accessible online such as the self-organizing map-based prediction of drug equivalence relationships (SPIDER) [106] and the Antibiotic'ome [108].

4.4 Binding Pose and Activity Predictions

If a broad set of structurally very similar molecules and their biological activity in a certain assay is well described, quantitative structure-activity relationship (QSAR) models can be calculated. Schmidt et al. used the information on 69 sesquiterpene lactone structures and their antitrypanosomal activities to generate a predictive model. The query was able to predict correctly furanoheliangolides with highly potent antitrypanosomal *in vitro* activity out of a virtual sesquiterpene database [150].

Molecular docking in combination with molecular dynamic simulations but also pharmacophore alignments have been demonstrated to accurately predict the binding mode of natural products to their respective targets offering valuable support for the understanding of bioactivities on a molecular level. Rollinger et al. used a combination of molecular docking and pharmacophore-based virtual screening to identify experimentally novel inhibitors of the human rhinovirus (HRV) capsid binders and to give insights into the interaction of natural product-derived inhibitors in the binding pocket. They proposed an eight-feature pharmacophore necessary for the identified ligands interacting in the binding site in addition to their fitting and binding mechanism into the highly lipophilic pocket [96].

The structure and function of membrane-bound GPCRs is still not well understood due to their difficult crystallization. Binding mechanisms of their ligands are nevertheless crucial, since approximately one third of all drugs target these proteins. After identifying several alkaloids as 5-HT_{2C} receptor ligands with a combined virtual and experimental screening, Peng et al. used a homology model to predict the interaction pattern of the ligands. Molecular docking and molecular dynamics suggested key interactions such as a conserved salt bridge and π stacking [101].

5 T: Threats of Cheminformatics in Natural Product Research

At first glance, the broad use of natural products in the field of cheminformatics should not lead to overestimated perceptions. As outlined in Sect. 3 weaknesses are pervasive and experiments are mandatory for confirmation of results. However, commonly, this is not the case for binding mode predictions, which frequently are reported without any proof of correctness.

Molecular target prediction tools are similarly hard to evaluate experimentally and natural product researchers should scrutinize retrieved predictions with healthy skepticism. The biomedical data for natural products is comparatively small when compared to other molecule classes. Therefore, it must be assumed that they are generally underrepresented in generation and validation of computational models. This might not only be the case for target prediction but also for the estimation of lipophilicity, conformer generation, assay interference prediction, molecular docking force-field adjustment, and other tasks.

In silico models must follow scripted instructions and generate only predictions. Flexibility, dynamics, entropic issues along with many more aspects can only be approached with extensive computational efforts. Virtual screening experiments still produce many false-positive virtual hits and incorrect or distorted results. Accordingly, predictions without any solid and unbiased experimental validation are not able to stand any test of scientific meaningfulness and therefore have to be regarded as "preliminary." On the other hand, even if experimentally validated, the probability of not being able to gain access to information of experimentally proven wrong hypothesis/models is very high. This not only refers to models that failed a proof of concept but also to test data of compounds showing no activity on a specific target. With special regard to the correct feeding and training of prediction tools with structural data covering a broad range of activity, ideally from inactive compounds to highly potent ones, learning from previous mistakes and non-working hypotheses would be extremely valuable. The fact that so many successful projects have been reported disguises the fact that other projects failed.

The availability of natural products in sufficient purity from commercial suppliers or obtaining these by isolation from a suitable natural source can be very costly or time-intensive. The natural starting material should be accessible and legally available for collection/acquisition considering issues on bioprospecting, intellectual property rights, and transfer of natural material to the outside its country of origin (Nagoya protocol, [155]). Also reliable reports on the natural product isolation procedure as well as compound structure elucidation parameters and the description of relevant physicochemical properties should be accessible for a target-oriented re-isolation and identification using mass spectrometry-based dereplication.

Special attention should be devoted to broadly distributed PAINS motifs in natural products such as catechols, hydroquinones, epoxides, peroxide bridges, and phenolic Mannich bases. Other concerns are solubility problems and compound aggregation. However, it might be inadvisable to generate a naive black-box application of PAINS and general drug-likeness filters [156] without looking beyond these parameters.

6 Conclusion

The process of small-molecule drug discovery can be described as being deterministic and nonlinear (e.g., activity cliffs) resembling a chaotic system. This is particularly true for drug discovery from natural products, where researchers are

confronted along with nonlinear behavior, serendipitous events, errors, and incompleteness also from biological variance, complex multicomponent mixture interactions, and frequent assay interferences. The current challenge of medicinal chemists is to choose which of the possible 10^{60} drug-like molecules should be synthesized and tested [18]. Considering the historical impact of natural products on the pharmaceutical arsenal and their infinite (however, incompletely known) diversity, secondary metabolites have already been synthesized by the most trained chemist on earth and thus are hidden gems designed to have key functions. In natural product research, the application of cheminformatics-based strategies is limited to already structurally disclosed molecules; accordingly, their potentially very large impact relies on properly performed and trustworthy chemical studies on natural resources and their constituents and their documentation and dissemination.

The technological advances and experimental exploration of the last centuries, in particular, have afforded the opportunity of accessing enormous amounts of data. Selecting the appropriate computational tools for handling these data and for addressing the research question is a key step but still requires a healthy skepticism and an unbiased attitude.

The Nobel Laureate Rolf Zinkernagel once made a piercing summary of different research strategies and their chance for success [185]: Having no rationale and performing no experiments is cheap but will not lead to results. To start from a rationale, but renounce experimental work is another relatively cheap method, but similarly does not lead to results. Lots of experiments without any rationale may produce interesting and serendipitous results, but with a disproportionate effort and waste of capacities. To perform experimental studies with a rationale is without surprise the gold method with a good yield of results and appropriate expenses. The generation of this rationale assisted by the use of already available data and with modern computational techniques based on the combined expertise from natural product researchers and computational chemists harbors the key to successful drug discovery processes in the field of remedies from Mother Nature.

References

1. Morimoto M, Komai K (2000) Plant secondary metabolites as plant defense systems. *Recent Res Dev Phytochem* 4:99
2. Hadacek F (2002) Secondary metabolites as plant traits: current assessment and future perspectives. *Crit Rev Plant Sci* 21:273
3. Moghe GD, Last RL (2015) Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiol* 169:1512
4. Wöll S, Kim SH, Greten HJ, Efferth T (2013) Animal plant warfare and secondary metabolite evolution. *Nat Prod Bioprospect* 3:1
5. van Hattum H, Waldmann H (2014) Biology-oriented synthesis: harnessing the power of evolution. *J Am Chem Soc* 136:11853
6. Sukuru SC, Jenkins JL, Beckwith RE, Scheiber J, Bender A, Mikhailov D, Davies JW, Glick M (2009) Plate-based diversity selection based on empirical HTS data to enhance the number of hits and their chemical diversity. *J Biomol Screen* 14:690

7. Larsson J, Gottfries J, Muresan S, Backlund A (2007) ChemGPS-NP: tuned for navigation in biologically relevant chemical space. *J Nat Prod* 70:789
8. Ertl P, Schuffenhauer A (2008) Cheminformatics analysis of natural products: lessons from Nature inspiring the design of new drugs. In: Petersen F, Amstutz R (eds) *Natural compounds as drugs*, vol II. Birkhäuser, Basel, p 217
9. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, Waldmann H (2005) Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc Natl Acad Sci U S A* 102:17272
10. Camp D, Garavelas A, Campitelli M (2015) Analysis of physicochemical properties for drugs of natural origin. *J Nat Prod* 78:1370
11. Stratton CF, Newman DJ, Tan DS (2015) Cheminformatic comparison of approved drugs from natural product versus synthetic origins. *Bioorg Med Chem Lett* 25:4802
12. Wetzel S, Schuffenhauer A, Roggo S, Ertl P, Waldmann H (2007) Cheminformatic analysis of natural products and their chemical space. *Chimia* 61:355
13. Lopez-Vallejo F, Giulianotti MA, Houghten RA, Medina-Franco JL (2012) Expanding the medicinally relevant chemical space with compound libraries. *Drug Discov Today* 17:718
14. Rollinger JM, Langer T, Stuppner H (2006) Strategies for efficient lead structure discovery from natural products. *Curr Med Chem* 13:1491
15. Newman DJ, Cragg GM (2016) Natural products as sources of new drugs from 1981 to 2014. *J Nat Prod* 79:629
16. Rodrigues T, Reker D, Schneider P, Schneider G (2016) Counting on natural products for drug design. *Nat Chem* 8:531
17. Strohl WR (2000) The role of natural products in a modern drug discovery program. *Drug Discov Today* 5:39
18. Schneider G (2017) Automating drug discovery. *Nat Rev Drug Discov* 17:97
19. Keller TH, Shi P-Y, Wang Q-Y (2011) Anti-infectives: can cellular screening deliver? *Curr Opin Chem Biol* 15:529
20. Swinney DC, Anthony J (2011) How were new medicines discovered? *Nat Rev Drug Discov* 10:507
21. Harrison S, Lahue B, Peng Z, Donofrio A, Chang C, Glick M (2017) Extending “predict first” to the design-make-test cycle in small-molecule drug discovery. *Future Med Chem* 9:533
22. Weller MG (2012) A unifying review of bioassay-guided fractionation, effect-directed analysis and related techniques. *Sensors* 12:9181
23. Kaur K, Michael H, Arora S, Harkonen P, Kumar S (2005) In vitro bioactivity-guided fractionation and characterization of polyphenolic inhibitory fractions from *Acacia nilotica* (L.) Willd. ex Del. *J Ethnopharmacol* 99:353
24. Yang X, Summerhurst DK, Koval SF, Ficker C, Smith ML, Bernards MA (2001) Isolation of an antimicrobial compound from *Impatiens balsamina* L. using bioassay-guided fractionation. *Phytother Res* 15:676
25. Cardellina JH, Munro MHG, Fuller RW, Manfredi KP, McKee TC, Tischler M, Bokesch HR, Gustafson KR, Beutler JA, Boyd MR (1993) A chemical screening strategy for the dereplication and prioritization of HIV-inhibitory aqueous natural products extracts. *J Nat Prod* 56:1123
26. Bindseil KU, Jakupovic J, Wolf D, Lavayre J, Leboul J, van der Pyl D (2001) Pure compound libraries; a new perspective for natural product based drug discovery. *Drug Discov Today* 6:840
27. Rao KV (1993) Taxol and related taxanes. I. Taxanes of *Taxus brevifolia* bark. *Pharm Res* 10:521
28. Chen Y, de Bruyn Kops C, Kirchmair J (2017) Data resources for the computer-guided discovery of bioactive natural products. *J Chem Inf Model* 57:2099
29. Reker D, Perna AM, Rodrigues T, Schneider P, Reutlinger M, Monch B, Koeberle A, Lamers C, Gabler M, Steinmetz H, Muller R, Schubert-Zsilavec M, Werz O, Schneider G (2014) Revealing the macromolecular targets of complex natural products. *Nat Chem* 6:1072

30. Kirchweger B, Rollinger JM (2018) Virtual screening for the discovery of active principles from natural products. In: Cechinel-Filho V (ed) *Natural products as source of molecules with therapeutic potential: research & development, challenges and perspectives*. Springer, Cham, pp 333–364
31. Rollinger JM, Langer T, Stuppner H (2006) Integrated in silico tools for exploiting the natural products' bioactivity. *Planta Med* 72:671
32. Rollinger JM, Quinn RJ (2015) In silico driven pharmacognosy: forth, back and reverse. *Planta Med* 81:427
33. Gasteiger J, Engel T (eds) (2006) *Chemoinformatics: a textbook*. Wiley-VCH, Weinheim
34. Rester U (2008) From virtuality to reality – virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Curr Opin Drug Discov Devel* 11:559
35. Rollinger JM, Wolber G (2011) Computational approaches for the discovery of natural lead structures. In: Tringali C (ed) *Bioactive compounds from natural sources*, 2nd edn. CRC Press, Boca Raton, FL, p 167
36. Schneider P, Schneider G (2017) De-orphaning the marine natural product (\pm)-marinopyrrole A by computational target prediction and biochemical validation. *Chem Commun* 53:2272
37. Rodrigues T, Sieglitz F, Somovilla VJ, Cal PM, Galione A, Corzana F, Bernardes GJ (2016) Unveiling (–)-englerin A as a modulator of L-type calcium channels. *Angew Chem Int Ed Eng* 55:11077
38. Kratz JM, Grienke U, Scheel O, Mann SA, Rollinger JM (2017) Natural products modulating the hERG channel: heartaches and hope. *Nat Prod Rep* 34:957
39. Lee ML, Schneider G (2001) Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries. *J Comb Chem* 3:284
40. Friedrich L, Rodrigues T, Neuhaus CS, Schneider P, Schneider G (2016) From complex natural products to simple synthetic mimetics by computational de novo design. *Angew Chem Int Ed Eng* 55:6789
41. Grisoni F, Merk D, Consonni V, Hiss JA, Tagliabue SG, Todeschini R, Schneider G (2018) Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Commun Chem* 1:44
42. Sichao W, Youyong L, Lei X, Dan L, Tingjun H (2013) Recent developments in computational prediction of hERG blockage. *Curr Top Med Chem* 13:1317
43. Scior T, Bernard P, Medina-Franco JL, Maggiora GM (2007) Large compound databases for structure-activity relationships studies in drug discovery. *Mini-Rev Med Chem* 7:851
44. Walters WP (2019) Virtual chemical libraries. *J Med Chem* 62:1116
45. Southan C (2018) Caveat usor: assessing differences between major chemistry databases. *ChemMedChem* 13:470
46. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40:D1100
47. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 42(D1):D1083
48. EMBL-EBI (2019) Homepage of the ChEMBL database, 2019. <https://www.ebi.ac.uk/chembl/>
49. NCBI (2019) PubChem, 2019. <https://pubchem.ncbi.nlm.nih.gov/>
50. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH (2016) PubChem substance and compound databases. *Nucleic Acids Res* 44(D1):D1202
51. Papadatos G, Davies M, Dedman N, Chambers J, Gaulton A, Siddle J, Koks R, Irvine SA, Pettersson J, Goncharoff N, Hersey A, Overington JP (2016) SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res* 44(D1):D1220
52. EMBL-EBI (2019) SureChEMBL, 2019. <https://www.surechembl.org/search/>

53. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 46(D1):D1074
54. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 34(Database issue):D668
55. DrugBank (2019) The DrugBank database, 2019. <https://www.drugbank.ca/>
56. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucleic Acids Res* 28:235
57. wwPDB consortium (2018) Protein data bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* 47(D1):D520
58. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 44(D1):D1045
59. Irwin JJ (2008) Community benchmarks for virtual screening. *J Comput Aided Mol Des* 22:193
60. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of Useful Decoys, Enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* 55:6582
61. Chan WK, Zhang H, Yang J, Brender JR, Hur J, Özgür A, Zhang Y (2015) GLASS: a comprehensive database for experimentally validated GPCR-ligand associations. *Bioinformatics* 31:3035
62. Excelra Knowledge Solutions (2019) GOSTAR database, 2019. <https://www.gostardb.com/gostar/index.jsp>
63. Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang Q-Y, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comp-Aided Mol Design* 25:533
64. OCHEM (2019) Online chemical database, 2019. <https://ochem.eu/home/show.do>
65. Li J, Liu J, Han L, Wang R, Nie W, Li Y, Liu Y, Liu Z, Zhao Z (2014) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* 31:405
66. Ahmed A, Dunbar JB Jr, Clark JJ, Smith RD, Carlson HA (2014) Recent improvements to Binding MOAD: a resource for protein–ligand binding affinities and structures. *Nucleic Acids Res* 43(D1):D465
67. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA (2005) Binding MOAD (mother of all databases). *Proteins Struct Funct Bioinf* 60:333
68. Frolkis A, Knox C, Lim E, Jewison T, Law V, Hau DD, Liu P, Gautam B, Ly S, Guo AC, Xia J, Liang Y, Shrivastava S, Wishart DS (2010) SMPDB: the small molecule pathway database. *Nucleic Acids Res* 38:D480
69. Jewison T, Su Y, Disfany FM, Liang Y, Knox C, Maciejewski A, Poelzer J, Huynh J, Zhou Y, Arndt D, Djoumbou Y, Liu Y, Deng L, Guo AC, Han B, Pon A, Wilson M, Rafatnia S, Liu P, Wishart DS (2014) SMPDB 2.0: big improvements to the small molecule pathway database. *Nucleic Acids Res* 42:D478
70. Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, Fu T, Zhang X, Cui X, Tu G, Zhang Y, Li S, Yang F, Sun Q, Qin C, Zeng X, Chen Z, Chen YZ, Zhu F (2017) Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res* 46:D1121
71. Swiss Institute of Bioinformatics (2019) Directory of computer-aided drug design tools, 2019. <https://www.click2drug.org/>

72. Kaserer T, Schuster D, Rollinger JM (2018) Chemoinformatics in natural product research. In: Engel T, Gasteiger J (eds) Applied chemoinformatics: achievements and future opportunities. Wiley-VCH, Weinheim, p 207
73. Harvey AL, Edrada-Ebel R, Quinn RJ (2015) The re-emergence of natural products for drug discovery in the genomics era. *Nat Rev Drug Discov* 14:111
74. Blunt J, Munro M, Upjohn M (2012) The role of databases in marine natural products research. In: Fattorusso E, Gerwick WH, Tagliatalata-Scafati O (eds) Natural compounds as drugs, 2nd edn. Springer, Dordrecht, p 389
75. Mohamed A, Nguyen CH, Mamitsuka H (2016) Current status and prospects of computational resources for natural product dereplication: a review. *Brief Bioinform* 17:309
76. Ma DL, Chan DSH, Leung CH (2011) Molecular docking for virtual screening of natural product databases. *Chem Sci* 2:1656
77. Blunt JW, Munro MHG (2014) Is there an ideal database for natural products research? In: Osbourn A, Goss RJ, Carter GT (eds) Natural products. Wiley-VCH, Weinheim, p 413
78. CRC Press, Taylor & Francis Group (2019) Dictionary of natural products 27.2., 2019. <http://dnp.chemnetbase.com>
79. Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X (2013) Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One* 8:e62839
80. Chen CY-C (2011) TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One* 6:e15939
81. Choi H, Cho SY, Pak HJ, Kim Y, Choi J-Y, Lee YJ, Gong BH, Kang YS, Han T, Choi G, Cho Y, Lee S, Ryoo D, Park H (2017) NPCARE: database of natural products and fractional extracts for cancer regulation. *J Cheminf* 9:2
82. Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2017) NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci Rep* 7:7215
83. Banerjee P, Erehman J, Gohlke B-O, Wilhelm T, Preissner R, Dunkel M (2015) Super Natural II—a database of natural products. *Nucleic Acids Res* 43:D935
84. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* 52:1757
85. Sigma-Aldrich (2019) Aldrich market select, 2019. <https://www.sigmaaldrich.com/chemistry/chemistry-services/aldrich-market-select.html>
86. Ertl P, Roggo S, Schuffenhauer A (2008) Natural product-likeness score and its application for prioritization of compound libraries. *J Chem Inf Model* 48:68
87. Jayaseelan KV, Moreno P, Truskowski A, Ertl P, Steinbeck C (2012) Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinf* 13:106
88. RDKIT Open-source cheminformatics software (2019) RDKIT version 2017.09.3, 2019. <http://www.rdkit.org/>
89. Yu MJ (2011) Natural product-like virtual libraries: recursive atom-based enumeration. *J Chem Inf Model* 51:541
90. Jayaseelan KV, Steinbeck C (2014) Building blocks for automated elucidation of metabolites: natural product-likeness for candidate ranking. *BMC Bioinf* 15:234
91. Zaid H, Raiyn J, Nasser A, Saad B, Rayan A (2010) Physicochemical properties of natural based products versus synthetic chemicals. *Open Nutraceuticals J* 3:194
92. Chen Y, Stork C, Hirte S, Kirchmair J (2019) NP-Scout: machine learning approach for the quantification and visualization of the natural product-likeness of small molecules. *Biomol Ther* 9:43
93. Harvey AL (2000) Natural products in drug discovery. *Drug Discov Today* 13:894
94. Tetko IV, Engkvist O, Koch U, Reymond J-L, Chen H (2016) BIGCHEM: challenges and opportunities for big data analysis in chemistry. *Mol Inf* 35:615
95. Rollinger JM (2009) Accessing target information by virtual parallel screening – The impact on natural product research. *Phytochem Lett* 2:53

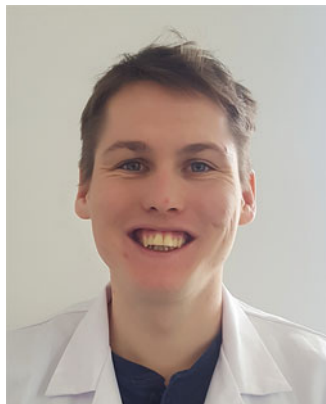
96. Rollinger JM, Steindl TM, Schuster D, Kirchmair J, Anrain K, Ellmerer EP, Langer T, Stuppner H, Wutzler P, Schmidtke M (2008) Structure-based virtual screening for the discovery of natural inhibitors for human rhinovirus coat protein. *J Med Chem* 51:842
97. Grienke U, Schmidtke M, Kirchmair J, Pfarr K, Wutzler P, Dürrwald R, Wolber G, Liedl KR, Stuppner H, Rollinger JM (2010) Antiviral potential and molecular insight into neuraminidase inhibiting diarylheptanoids from *Alpinia katsumadai*. *J Med Chem* 53:778
98. Leláková V, Šmejkal K, Jakubczyk K, Veselý O, Landa P, Václavík J, Bobál P, Pížová H, Temml V, Steinacher T, Schuster D, Granica S, Hanáková Z, Hošek J (2019) Parallel in vitro and in silico investigations into anti-inflammatory effects of non-prenylated stilbenoids. *Food Chem* 285:431
99. Rollinger JM, Kratschmar DV, Schuster D, Pfisterer PH, Gumy C, Aubry EM, Brandstotter S, Stuppner H, Wolber G, Odermatt A (2010) 11β -Hydroxysteroid dehydrogenase 1 inhibiting constituents from *Eriobotrya japonica* revealed by bioactivity-guided isolation and computational approaches. *Bioorg Med Chem* 18:1507
100. Fu W, Chen L, Wang Z, Zhao C, Chen G, Liu X, Dai Y, Cai Y, Li C, Zhou J, Liang G (2016) Determination of the binding mode for anti-inflammatory natural product xanthohumol with myeloid differentiation protein 2. *Drug Des Dev Ther* 10:455
101. Peng Y, Zhao S, Wu Y, Cao H, Xu Y, Liu X, Shui W, Cheng J, Zhao S, Shen L, Ma J, Quinn RJ, Stevens RC, Zhong G, Liu Z-J (2018) Identification of natural products as novel ligands for the human 5-HT_{2C} receptor. *Biophys Rep* 4:50
102. Petersen RK, Christensen KB, Assimopoulou AN, Fretté X, Papageorgiou VP, Kristiansen K, Kouskoumvekaki I (2011) Pharmacophore-driven identification of PPAR γ agonists from natural sources. *J Comput Aided Mol Des* 25:107
103. Zaheer-ul-Haq ZU, Wellenzohn B, Liedl KR, Rode BM (2003) Molecular docking studies of natural cholinesterase-inhibiting steroidal alkaloids from *Sarcococca saligna*. *J Med Chem* 46:5087
104. Atanasov AG, Wang JN, Gu SP, Bu J, Kramer MP, Baumgartner L, Fakhrudin N, Ladurner A, Malainer C, Vuorinen A, Noha SM, Schwaiger S, Rollinger JM, Schuster D, Stuppner H, Dirsch VM, Heiss EH (2013) Honokiol: a non-adipogenic PPAR γ agonist from Nature. *Biochim Biophys Acta* 1830:4813
105. Mulholland K, Wu C (2016) Binding of telomestatin to a telomeric G-quadruplex DNA probed by all-atom molecular dynamics simulations with explicit solvent. *J Chem Inf Model* 56:2093
106. ETH Zürich (2019) SPiDER Target Prediction Software, 2019. <http://modlab.cadd.ethz.ch/software/spider/>
107. Petra S, Gisbert S (2017) A computational method for unveiling the target promiscuity of pharmacologically active compounds. *Angew Chem Int Ed* 56:11520
108. Nathan Magarvey Lab (2019) Antibiotic'ome, 2019. <https://magarveylab.ca/antibioticome/#/search>
109. Johnston CW, Skinnider MA, Dejong CA, Rees PN, Chen GM, Walker CG, French S, Brown ED, Bérdy J, Liu DY, Magarvey NA (2016) Assembly and clustering of natural antibiotics guides target identification. *Nat Chem Biol* 12:233
110. Grienke U, Kaserer T, Pfluger F, Mair CE, Langer T, Schuster D, Rollinger JM (2015) Accessing biological actions of *Ganoderma* secondary metabolites by in silico profiling. *Phytochemistry* 114:114
111. Rollinger JM, Schuster D, Danzl B, Schwaiger S, Markt P, Schmidtke M, Gertsch J, Raduner S, Wolber G, Langer T, Stuppner H (2009) In silico target fishing for rationalized ligand discovery exemplified on constituents of *Ruta graveolens*. *Planta Med* 75:195
112. Gong J, Sun P, Jiang N, Riccio R, Lauro G, Bifulco G, Li T-J, Gerwick WH, Zhang W (2014) New steroids with a rearranged skeleton as (h)P300 inhibitors from the sponge *Theonella swinhoei*. *Org Lett* 16:2224
113. Di Micco S, Pulvirenti L, Bruno I, Terracciano S, Russo A, Vaccaro MC, Ruggiero D, Muccilli V, Cardullo N, Tringali C, Riccio R, Bifulco G (2018) Identification by inverse

- virtual screening of magnolol-based scaffold as new tankyrase-2 inhibitors. *Bioorg Med Chem* 26:3953
114. Rollinger JM, Hornick A, Langer T, Stuppner H, Prast H (2004) Acetylcholinesterase inhibitory activity of scopolin and scopoletin discovered by virtual screening of natural products. *J Med Chem* 47:6248
 115. Rollinger JM, Haupt S, Stuppner H, Langer T (2004) Combining ethnopharmacology and virtual screening for lead structure discovery: COX-inhibitors as application example. *J Chem Inf Comput Sci* 44:480
 116. Rollinger JM, Bodensieck A, Seger C, Ellmerer EP, Bauer R, Langer T, Stuppner H (2005) Discovering COX-inhibiting constituents of *Morus* root bark: activity-guided versus computer-aided methods. *Planta Med* 71:399
 117. Kratz JM, Mair CE, Oetl SK, Saxena P, Scheel O, Schuster D, Hering S, Rollinger JM (2016) hERG channel blocking ipecac alkaloids identified by combined in silico – in vitro screening. *Planta Med* 82:1009
 118. Kratz JM, Schuster D, Edtbauer M, Saxena P, Mair CE, Kirchebner J, Matuszczak B, Baburin I, Hering S, Rollinger JM (2014) Experimentally validated hERG pharmacophore models as cardiotoxicity prediction tools. *J Chem Inf Model* 54:2887
 119. Grienke U, Mihaly-Bison J, Schuster D, Afonyushkin T, Binder M, Guan SH, Cheng CR, Wolber G, Stuppner H, Guo DA, Bochkov VN, Rollinger JM (2011) Pharmacophore-based discovery of FXR-agonists. Part II: identification of bioactive triterpenes from *Ganoderma lucidum*. *Bioorg Med Chem* 19:6779
 120. Schuster D, Markt P, Grienke U, Mihaly-Bison J, Binder M, Noha SM, Rollinger JM, Stuppner H, Bochkov VN, Wolber G (2011) Pharmacophore-based discovery of FXR agonists. Part I: model development and experimental validation. *Bioorg Med Chem* 1:7168
 121. Noha SM, Jazzar B, Kuehl S, Rollinger JM, Stuppner H, Schaible AM, Werz O, Wolber G, Schuster D (2012) Pharmacophore-based discovery of a novel cytosolic phospholipase A(2) α inhibitor. *Bioorg Med Chem Lett* 22:1202
 122. Waltenberger B, Wiechmann K, Bauer J, Markt P, Noha SM, Wolber G, Rollinger JM, Werz O, Schuster D, Stuppner H (2011) Pharmacophore modeling and virtual screening for novel acidic inhibitors of microsomal prostaglandin E₂ synthase-1 (mPGES-1). *J Med Chem* 54:3163
 123. Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO (2005) Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J Med Chem* 48:2534
 124. Choi J, Kim H-J, Jin X, Lim H, Kim S, Roh I-S, Kang H-E, No KT, Sohn H-J (2018) Application of the fragment molecular orbital method to discover novel natural products for prion disease. *Sci Rep* 8:13063
 125. Kirchweber B, Kratz JM, Ladurner A, Grienke U, Langer T, Dirsch VM, Rollinger JM (2018) In silico workflow for the identification of natural products targeting GPBAR1. *Front Chem* 6:242
 126. Vuorinen A, Seibert J, Papageorgiou VP, Rollinger JM, Odermatt A, Schuster D, Assimpoulou AN (2015) *Pistacia lentiscus* oleoresin: virtual screening and identification of masticadienonic and isomasticadienonic acids as inhibitors of 11 β -hydroxysteroid dehydrogenase I. *Planta Med* 81:525
 127. Fakhrudin N, Ladurner A, Atanasov AG, Heiss EH, Baumgartner L, Markt P, Schuster D, Ellmerer EP, Wolber G, Rollinger JM, Stuppner H, Dirsch VM (2010) Computer-aided discovery, validation, and mechanistic characterization of novel neolignan activators of peroxisome proliferator-activated receptor gamma. *Mol Pharmacol* 77:559
 128. Duwensee K, Schwaiger S, Tancevski I, Eller K, van Eck M, Markt P, Linder T, Stanzl U, Ritsch A, Patsch JR, Schuster D, Stuppner H, Bernhard D, Eller P (2011) Leoligin, the major lignan from Edelweiss, activates cholesteryl ester transfer protein. *Atherosclerosis* 219:109

129. Maindola P, Jamal S, Grover A (2015) Cheminformatics based machine learning models for AMA1-RON2 abrogators for inhibiting *Plasmodium falciparum* erythrocyte invasion. *Mol Inf* 34:655
130. Herrmann FC, Lenz M, Jose J, Kaiser M, Brun R, Schmidt TJ (2015) In silico identification and in vitro activity of novel natural inhibitors of *Trypanosoma brucei* glyceraldehyde-3-phosphate-dehydrogenase. *Molecules* 20:16154
131. Karhu E, Isojärvi J, Vuorela P, Hanski L, Fallarero A (2017) Identification of privileged antichlamydial natural products by a ligand-based strategy. *J Nat Prod* 80:2602
132. Diao Y, Jiang J, Zhang S, Li S, Shan L, Huang J, Zhang W, Li H (2018) Discovery of natural products as novel and potent FXR antagonists by virtual screening. *Front Chem* 6:140
133. Grienke U, Braun H, Seidel N, Kirchmair J, Richter M, Krumbholz A, von Grafenstein S, Liedl KR, Schmidtke M, Rollinger JM (2014) Computer-guided approach to access the anti-influenza activity of licorice constituents. *J Nat Prod* 77:563
134. Noha SM, Atanasov AG, Schuster D, Markt P, Fakhruddin N, Heiss EH, Schrammel O, Rollinger JM, Stuppner H, Dirsch VM, Wolber G (2011) Discovery of a novel IKK- β inhibitor by ligand-based virtual screening techniques. *Bioorg Med Chem Lett* 21:577
135. Bauer J, Waltenberger B, Noha SM, Schuster D, Rollinger JM, Boustie J, Chollet M, Stuppner H, Werz O (2012) Discovery of depsides and depsidones from lichen as potent inhibitors of microsomal prostaglandin E₂ synthase-1 using pharmacophore models. *ChemMedChem* 7:2077
136. Costa G, Rocca R, Corona A, Grandi N, Moraca F, Romeo I, Talarico C, Gagliardi MG, Ambrosio FA, Ortuso F, Alcaro S, Distinto S, Maccioni E, Tramontano E, Artese A (2019) Novel natural non-nucleoside inhibitors of HIV-1 reverse transcriptase identified by shape- and structure-based virtual screening techniques. *Eur J Med Chem* 161:1
137. Carrasco MP, Gut J, Rodrigues T, Ribeiro MHL, Lopes F, Rosenthal PJ, Moreira R, dos Santos DJ (2013) Exploring the molecular basis of Qo bc1 complex inhibitors activity to find novel antimalarials hits. *Mol Inf* 32:659
138. Rodrigues T, Ressurreição AS, da Cruz FP, Albuquerque IS, Gut J, Carrasco MP, Gonçalves D, Guedes RC, dos Santos DJ, Mota MM, Rosenthal PJ, Moreira R, Prudêncio M, Lopes F (2013) Flavones as isosteres of 4(1*H*)-quinolones: discovery of ligand efficient and dual stage antimalarial lead compounds. *Eur J Med Chem* 69:872
139. Ikram NKK, Durrant JD, Muchtaridi M, Zalaludin AS, Purwitasari N, Mohamed N, Rahim ASA, Lam CK, Normi YM, Rahman NA, Amaro RE, Wahab HA (2015) A virtual screening approach for identifying plants with anti H5N1 neuraminidase activity. *J Chem Inf Model* 55:308
140. Ou T, Hou X, Guan S, Dai J, Han W, Li R, Wang W, Qu X, Zhang M (2016) Targeting AMPK signalling pathway with natural medicines for atherosclerosis therapy: an integration of in silico screening and in vitro assay. *Nat Prod Res* 30:1240
141. Su H, Yan J, Xu J, Fan XZ, Sun XL, Chen KY (2015) Stepwise high-throughput virtual screening of Rho kinase inhibitors from natural product library and potential therapeutics for pulmonary hypertension. *Pharm Biol* 53:1201
142. Schuster D, Kern L, Hristozov DP, Terfloth L, Bienfait B, Laggner C, Kirchmair J, Grienke U, Wolber G, Langer T (2010) Applications of integrated data mining methods to exploring natural product space for acetylcholinesterase inhibitors. *Comb Chem High Throughput Screen* 13:54
143. Cozza G, Bonvini P, Zorzi E, Poletto G, Pagano MA, Sarno S, Donella-Deana A, Zagotto G, Rosolen A, Pinna LA, Meggio F, Moro S (2006) Identification of ellagic acid as potent inhibitor of protein kinase CK2: a successful example of a virtual screening application. *J Med Chem* 49:2363
144. Salam NK, Huang TH, Kota BP, Kim MS, Li Y, Hibbs DE (2008) Novel PPAR- γ agonists identified from a natural product library: a virtual screening, induced-fit docking and biological assay study. *Chem Biol Drug Des* 71:57

145. Karaman B, Alhalabi Z, Swyter S, Mihigo SO, Andrae-Marobela K, Jung M, Sippl W, Ntie-Kang F (2018) Identification of bichalcones as sirtuin inhibitors by virtual screening and in vitro testing. *Molecules* 23:416
146. Ke Z, Su Z, Zhang X, Cao Z, Ding Y, Cao L, Ding G, Wang Z, Liu H, Xiao W (2017) Discovery of a potent angiotensin converting enzyme inhibitor via virtual screening. *Bioorg Med Chem Lett* 27:3688
147. Chan DS-H, Lee H-M, Yang F, Che C-M, Wong CCL, Abagyan R, Leung C-H, Ma DL (2010) Structure-based discovery of natural-product-like TNF- α inhibitors. *Angew Chem Int Ed* 49:2860
148. Ma DL, Chan DS, Fu WC, He HZ, Yang H, Yan SC, Leung CH (2012) Discovery of a natural product-like c-myc G-quadruplex DNA groove-binder by molecular docking. *PLoS One* 7: e43278
149. Liu Y, Huang L, Ye H, Lv X (2016) Combined QSAR-based virtual screening and fluorescence binding assay to identify natural product mediators of interferon regulatory factor 7 (IRF-7) in pulmonary infection. *SAR QSAR Environ Res* 27:967
150. Schmidt TJ, Da Costa FB, Lopes NP, Kaiser M, Brun R (2014) In silico prediction and experimental evaluation of furanoheliangolide sesquiterpene lactones as potent agents against *Trypanosoma brucei rhodesiense*. *Antimicrob Agents Chemother* 58:325
151. Rupp M, Schroeter T, Steri R, Zettl H, Proschak E, Hansen K, Rau O, Schwarz O, Muller-Kuhrt L, Schubert-Zsilavecz M, Muller KR, Schneider G (2010) From machine learning to natural product derivatives that selectively activate transcription factor PPAR γ . *ChemMedChem* 5:191
152. Schuster D, Wolber G (2010) Identification of bioactive natural products by pharmacophore-based virtual screening. *Curr Pharm Des* 16:1666
153. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem* 45:2213
154. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, O'Meara MJ, Che T, Algaa E, Tolmacheva K, Tolmachev AA, Shoichet BK, Roth BL, Irwin JJ (2019) Ultra-large library docking for discovering new chemotypes. *Nature* 566:224
155. Matthias B, Clare H (2011) The Nagoya protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the Convention on Biological Diversity. *Rev Eur Commun Int Environ Law* 20:47
156. Baell JB (2016) Feeling Nature's PAINS: natural products, natural product drugs, and pan assay interference compounds (PAINS). *J Nat Prod* 79:616
157. Scior T, Bender A, Tresadern G, Medina-Franco JL, Martínez-Mayorga K, Langer T, Cuanalo-Contreras K, Agrafiotis DK (2012) Recognizing pitfalls in virtual screening: a critical review. *J Chem Inf Model* 52:867
158. Grienke U, Mair CE, Kirchmair J, Schmidtke M, Rollinger JM (2018) Discovery of bioactive natural products for the treatment of acute respiratory infections – an integrated approach. *Planta Med* 84:684
159. Chen Y, Garcia de Lomana M, Friedrich N-O, Kirchmair J (2018) Characterization of the chemical space of known and readily obtainable natural products. *J Chem Inf Model* 58:1518
160. Feher M, Schmidt JM (2003) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J Chem Inf Comput Sci* 43:218
161. Clemons PA, Bodycombe NE, Carrinski HA, Wilson JA, Shamji AF, Wagner BK, Koehler AN, Schreiber SL (2010) Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc Natl Acad Sci U S A* 107:18787
162. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* 45:2615
163. Mora CA, Halter JG, Adler C, Hund A, Anders H, Yu K, Stark WJ (2016) Application of the *Prunus* spp. cyanide seed defense system onto wheat: reduced insect feeding and field growth tests. *J Agric Food Chem* 64:3501

164. Zhou M, Zhang R-H, Wang M, Xu G-B, Liao S-G (2017) Prodrugs of triterpenoids and their derivatives. *Eur J Med Chem* 131:222
165. CDK Knime Nodepit (2019) Sugar Remover, 2019. <https://nodepit.com/node/org.openscience.cdk.knime.nodes.sugarremover.SugarRemoverNodeFactory>
166. Nivedha AK, Thieker DF, Makeneni S, Hu H, Woods RJ (2016) Vina-Carb: improving glycosidic angles during carbohydrate docking. *J Chem Theory Comput* 12:892
167. Clostre F (1999) *Ginkgo biloba* extract (EGb 761). State of knowledge in the dawn of the year 2000. *Ann Pharm Fr* 57(Suppl 1):1S8
168. Xu R, Zhao W, Xu J, Shao B, Qin G (1996) Studies on bioactive saponins from Chinese medicinal plants. *Adv Exp Med Biol* 404:371
169. Cesarone MR, Ricci A, Di Renzo A, Belcaro G, Dugall M (2004) Efficacy of topical treatment with aescin + essential phospholipids gel on capillary fragility. *Angiology* 55(Suppl 1):S23
170. Brandley BK, Schnaar RL (1986) Cell-surface carbohydrates in cell recognition and response. *J Leukoc Biol* 40:97
171. McBride A, Ghilagaber S, Nikolaev A, Hardie DG (2009) The glycogen-binding domain on the AMPK beta subunit allows the kinase to act as a glycogen sensor. *Cell Metab* 9:23
172. Kato K, Ishiwa A (2015) The role of carbohydrates in infection strategies of enteric pathogens. *Trop Med Health* 43:41
173. Lorent JH, Quetin-Leclercq J, Mingeot-Leclercq MP (2014) The amphiphilic nature of saponins and their effects on artificial and biological membranes and potential consequences for red blood and cancer cells. *Org Biomol Chem* 12:8803
174. Barbič M, Willer EA, Rothenhöfer M, Heilmann J, Fürst R, Jürgenliemk G (2013) Spirostanol saponins and esculin from *Ruscus Rhizoma* reduce the thrombin-induced hyperpermeability of endothelial cells. *Phytochemistry* 90:106
175. Sottriffer C (2018) Docking of covalent ligands: challenges and approaches. *Mol Inf* 37:1800062
176. Hardman TC, Dubrey SW (2011) Development and potential role of type-2 sodium-glucose transporter inhibitors for management of type 2 diabetes. *Diabetes Ther* 2:133
177. Gribble GW (1998) Naturally occurring organohalogen compounds. *Acc Chem Res* 31:141
178. Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG (2017) Retrospective analysis of natural products provides insights for future discovery trends. *Proc Natl Acad Sci U S A* 114:5601
179. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 46:3
180. Vuorinen A, Engeli R, Meyer A, Bachmann F, Griesser UJ, Schuster D, Odermatt A (2014) Ligand-based pharmacophore modeling and virtual screening for the discovery of novel 17 β -hydroxysteroid dehydrogenase 2 inhibitors. *J Med Chem* 57:5995
181. Vuorinen A, Nashev LG, Odermatt A, Rollinger JM, Schuster D (2014) Pharmacophore model refinement for 11 β -hydroxysteroid dehydrogenase inhibitors: search for modulators of intracellular glucocorticoid concentrations. *Mol Inf* 33:15
182. Trivedi PJ, Hirschfield GM, Gershwin ME (2016) Obeticholic acid for the treatment of primary biliary cirrhosis. *Expert Rev Clin Pharmacol* 9:13
183. Bozorgi M, Memariani Z, Mobli M, Salehi Surmaghi MH, Shams-Ardekani MR, Rahimi R (2013) Five *Pistacia* species (*P. vera*, *P. atlantica*, *P. terebinthus*, *P. khinjuk*, and *P. lentiscus*): a review of their traditional uses, phytochemistry, and pharmacology. *Sci World J D* 2013:219815
184. Georgiadis I, Karatzas T, Korou LM, Agrogiannis G, Vlachos IS, Pantopoulou A, Tzanetakou IP, Katsilambros N, Perrea DN (2014) Evaluation of Chios mastic gum on lipid and glucose metabolism in diabetic mice. *J Med Food* 17:393
185. Zinkernagel RM (1997) Lecture "Immunität gegen Viren" presented at the Symposium "Pharmazie. Die Wissenschaft vom Arzneimittel" of the German and Swiss Pharmaceutical Societies, Zürich, Switzerland, October 02–05



Benjamin Kirchweger is a doctoral student in the “Phytochemistry and Biodiscovery” group at the Department of Pharmacognosy, Faculty of Life Science, University of Vienna, Vienna, Austria; he studied Pharmacy, received his master’s degree in 2017 and started his doctoral studies under the supervision of Judith M. Rollinger. His main research interest is the practical application of computational techniques in natural product drug discovery, with a special focus on metabolic syndrome. He aims to streamline results from *in vivo* small animal models to *in vitro* and *in silico* models to discover and evaluate natural products affecting conserved metabolic pathways, such as bile acid metabolism, sirtuins, AMPK, and Nrf2. He used successfully a ligand-based virtual screening approach to identify new natural product activators of the G-protein-coupled bile acid receptor 1. Benjamin Kirchweger currently holds a position as university assistant at the University of Vienna.



Judith Maria Rollinger is a professor of Pharmacognosy/Pharmaceutical Biology and head of “Phytochemistry and Biodiscovery” at the Department of Pharmacognosy, Faculty of Life Science, University of Vienna, Austria. After obtaining her PhD degree in Pharmacognosy of the University of Innsbruck/Austria dealing with crystal polymorphism (1999), she extended her studies to the fields of phytochemistry, ethnopharmacology, and molecular modeling. Engaged as senior scientist of the software company Inte:Ligand, Austria, she implemented computational approaches in natural product science. In 2007, she received the “*venia legendi*” as professor of Pharmacognosy as a result of her habilitation thesis “The Search for Bioactive Natural Products.” Prof. Rollinger is project leader in various national and international projects and has received several awards in her field. She was appointed full professor of Pharmacognosy/Pharmaceutical Biology at her present institution in 2014. She is vice-president (since 2018) of the “Society of Medicinal Plants and Natural Product Research,” and an editor of *Planta Medica* (since 2016). Her research focuses on the interdisciplinary field of integrating computational techniques in pharmacognostic research as strategy for the discovery of natural lead structures for treating viral infections, metabolic syndrome, and inflammation. Publications resulting from her research have appeared in highly ranked international journals (>100) and as book contributions and patents.