

Cheminformatics in Natural Product-Based Drug Discovery

Cumulative Dissertation

with the aim of achieving the degree

Doctor rerum naturalium (Dr. rer. nat.)

at the Faculty of Mathematics, Informatics and Natural Sciences,

Department of Informatics,

Universität Hamburg

submitted by

Ya Chen

born in Zhengzhou, China

Hamburg, 2020

The presented thesis was prepared from October 2016 till August 2020 under the supervision of Dr. Johannes Kirchmair at the Department of Informatics, Universität Hamburg.

1. Reviewer: Dr. Johannes Kirchmair
2. Reviewer: Prof. Dr. Gerhard Wolber

Date of thesis defense: 10.11.2020

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

08.09.2020

Datum

Jalhen

Unterschrift

Abstract

Natural products (NPs) remain the single most prolific source of inspiration for small-molecule drug discovery. Boosted by the increasing amount of data available on the chemical, biological, pharmacological and structural properties of NPs, computational approaches have become a mainstay in NP research. In silico methods are particularly useful as decision support tools, allowing experimentalists to focus their resources on the most promising directions. However, the current knowledge of the quantity, quality and relevance of the available data as well as of the scope and limitations of cheminformatics methods in NP-based drug discovery is limited.

The aims of this PhD thesis are hence to (i) develop a comprehensive understanding of the data that can be utilized for the advancement and application of in silico methods in the context of NP research, (ii) develop a new method able to identify NPs and NP-like compounds in large compound collections, in order to maximize the use of the available chemical data, and (iii) determine the capacity of a three-dimensional shape-based method to predict the macromolecular targets of complex small molecules such as complex NPs.

In the first part of this work a comprehensive perspective on the scope and limitations of in silico methods in NP-based drug discovery is presented. This is followed by an exhaustive review of a large number of virtual and physical NP libraries that are relevant to applications in cheminformatics, especially in virtual screening. One result of this work is a comprehensive, carefully curated virtual collection of 250k NPs. By overlaying this database with a large set of readily obtainable small organic compounds we are able, for the first time, to estimate the number of readily obtainable NPs, which is in the range of 25k (10% of the known NPs).

In the next phase of this PhD thesis, we conduct an in-depth analysis of the physicochemical and structural properties of the known NPs, the readily obtainable NPs, and individual NP libraries, as well as compare them with those of approved drugs. An in silico algorithm for removing sugars and sugar-like moieties from NPs and a rule-based approach for the identification of different NP classes are developed. This study shows that NPs are highly diverse. The majority of readily obtainable NPs are found to populate areas in chemical space that are of direct relevance to drug discovery. For several NP databases, a large number of compounds are identified which cover distinct areas in chemical space.

One important learning from our survey of compound collections is that NPs are often mixed with NP derivatives and analogs, as well as with synthetic compounds. In fact, substantial numbers of potentially valuable NPs are included in commercial compound collections with no mention of NPs or with no labels that would allow their easy identification. This prompts us to develop a machine

learning approach that enables the automated cherry-picking of NPs and NP-like compounds from large compound collections. The method is based on a random forest algorithm that obtains a high classification accuracy on holdout data. Moreover, we implement a method that allows the visualization of the areas in a molecule that contribute to the classification of a compound as either a NP or synthetic compound. The best-performing models are provided via a free web service.

The final part of this thesis is dedicated to what is currently one of the hottest research topics in cheminformatics, which is the prediction of the macromolecular targets of small organic compounds. NPs pose a particular challenge to such methods because of the scarcity of available bioactivity data on related compounds and the structural complexity of many NPs. The capacity of a three-dimensional shape-based approach is systematically explored to identify the biomacromolecular targets of structurally complex small molecules (including large and flexible NPs and macrocyclic compounds) based on their similarity to non-complex small molecules (i.e. more conventional, "drug-like" synthetic compounds). This approach obtains good success rates even for compounds that are clearly distinct in their structure from any of the ligands present in the knowledge base. Cases of complete failure are recorded only for a small number of targets. However, complex NPs prove to be challenging even with this robust approach.

Overall, this PhD thesis provides a wealth of new information and in-depth knowledge on the available data and cheminformatics methods relevant to natural products-based drug discovery. The study has resulted in accurate models that allow the automated identification and extraction of NPs and NP-like compounds from compound collections, and in a thoroughly validated, three-dimensional shape-based approach for identifying the targets for complex small molecules, especially for complex NPs.

Zusammenfassung

Naturstoffe stellen weiterhin die wichtigste Inspirationsquelle für die Entwicklung moderner Wirkstoffe dar. Mit der zunehmenden Verfügbarkeit experimenteller Daten über die chemischen, biologischen, pharmakologischen und strukturellen Eigenschaften von Naturstoffen konnten sich computergestützte Methoden als eine tragende Technologie in der Erforschung von Naturstoffen etablieren. Die theoretischen Ansätze erlauben es, die limitierten experimentellen Ressourcen in die vielversprechendsten Richtungen zu leiten. Das derzeitige Wissen über die Quantität, Qualität und Relevanz der verfügbaren experimentellen Daten, sowie die Anwendungsbereiche und Grenzen moderner chemieinformatischer Methoden im Bereich der naturstoffbasierten Arzneimittelentwicklung, sind jedoch begrenzt.

Die Ziele dieser Doktorarbeit sind daher (i) die Entwicklung eines umfassenden Verständnisses über die verfügbaren experimentellen Daten, welche für die Weiterentwicklung und Anwendung von computerbasierten Methoden im Kontext der Naturstoffforschung genutzt werden können, (ii) die Entwicklung einer computerbasierten Methode für die automatisierte Erkennung von Naturstoffen und naturstoffähnlichen Verbindungen in großen Moleküldatenbanken (mit dem Ziel die Nutzung der verfügbaren chemischen Daten zu maximieren), und (iii) die Erforschung der Kapazität shape-basierter Methoden, die Zielproteine strukturell komplexer Wirkstoffe, einschließlich Naturstoffe, vorherzusagen.

Im ersten Teil dieser Arbeit wird eine umfassende Analyse der Anwendungsbereiche und Grenzen moderner chemieinformatischer Methoden in der Naturstoffforschung präsentiert. Anschließend werden die verfügbaren und für die computergestützte Arzneistoffentwicklung relevanten Naturstoffdatenbanken umfassend analysiert. Ein wesentliches Resultat dieser Arbeit ist eine sorgfältig zusammengestellte, umfangreiche, virtuelle Strukturdatensammlung von 250,000 Naturstoffen. Diese Moleküldatenbank wird mit einem umfassenden Datensatz der weltweit verfügbaren Substanzen verglichen. Dadurch kann zum ersten Mal die Anzahl der Naturstoffe abgeschätzt werden, die zeitnahe für eine experimentelle Testung zugänglich sind. Es handelt sich hierbei um etwa 25,000 Substanzen (dies entspricht 10% aller bekannten Naturstoffe).

In der nächsten Phase dieser Doktorarbeit werden physikalisch-chemische und strukturelle Eigenschaften der bekannten Naturstoffe und der verfügbaren Naturstoffe mit jenen der zugelassenen Arzneistoffe verglichen. Im Rahmen dieser Studie werden ein computerbasierter Algorithmus zur Entfernung von Zuckern und zucker-ähnlichen Fragmenten aus Naturstoffen sowie ein regelbasierter Ansatz für die Identifizierung verschiedener Naturstoffklassen vorgestellt. Die Arbeit zeigt die strukturelle Vielfalt der bekannten Naturstoffe. Viele Naturstoffe ähneln in ihren physikalisch-chemischen Eigenschaften jenen der Arzneistoffe,

andere Naturstoffe wiederum unterscheiden sich in diesen Eigenschaften deutlich von Arzneistoffen und decken andere Bereiche des chemischen Raums ab.

Eine wichtige Erkenntnis aus dieser Doktorarbeit ist, dass Naturstoffe, deren Derivate und Analoga, und synthetische Verbindungen in virtuellen Substanzbibliotheken oft gemischt vorliegen und nicht entsprechend gekennzeichnet sind. Deshalb wird im Rahmen dieser Arbeit ein maschinelles Lernverfahren entwickelt, das automatisch Naturstoffe und naturstoffähnliche Substanzen in großen Substanzdatenbanken identifizieren kann. Die Methode basiert auf einem Random-Forest Algorithmus und erzielt eine hohe Klassifikationsgenauigkeit. Zudem wird eine Methode zur Visualisierung der Molekülbereiche, die maßgeblich zur Klassifizierung einer Verbindung als Naturstoff beziehungsweise als synthetische Verbindung beitragen, implementiert. Die besten Modelle sind über einen Web Service kostenlos für die Öffentlichkeit zugänglich.

Der letzte Teil der Arbeit widmet sich der computerbasierten Vorhersage der Zielproteine kleiner organischer Verbindungen, einem hochaktuellen Forschungsthema der Chemieinformatik. Naturstoffe stellen aufgrund ihrer oft hohen Komplexität und der Knappheit der verfügbaren Bioaktivitätsdaten über verwandte Verbindungen eine besonders große Herausforderung für solche Ansätze dar. Konkret wird ein Ansatz, der auf dem Vergleich der dreidimensionalen Strukturen von Molekülen basiert, untersucht, um Zielproteine strukturell komplexer Wirkstoffe (einschließlich großer, flexibler Naturstoffe und makrozyklischer Verbindungen) vorherzusagen. Die Vorhersage basiert auf der Ähnlichkeit der zu untersuchenden Substanzen zu strukturell einfachen Wirkstoffmolekülen (d.h. konventionellen, "Medikamenten-ähnlichen", synthetischen Verbindungen). Mit diesem Ansatz werden gute Erfolgsraten selbst für Substanzen, deren Struktur sich deutlich von allen Referenzsubstanzen abheben, erzielt. Nur in wenigen Fällen kann die Methode nicht zur Aufklärung der Zielproteine beitragen. Naturstoffe erweisen sich jedoch auch für diesen robusten Ansatz als besonders anspruchsvoll.

Zusammenfassend liefert diese Doktorarbeit eine Vielzahl neuer Erkenntnisse über die für die Naturstoffforschung relevanten Datenbanken und chemieinformatischen Methoden und trägt somit zu einem tiefgehenden Verständnis bei. Im Rahmen der Arbeit werden genaue Modelle für die automatische Identifizierung und Extraktion von Naturstoffen und Naturstoff-ähnlichen Verbindungen aus Substanzbibliotheken entwickelt. Zudem wird ein gründlich validierter Ansatz, basierend auf dem Vergleich dreidimensionaler Moleküloberflächen, zur Identifizierung der Zielproteine strukturell komplexer Wirkstoffe (insbesondere komplexer Naturstoffe) erforscht.

Acknowledgements

First and foremost, I would like to thank my supervisor, Ass.-Prof. Dr. Johannes Kirchmair, for giving me the opportunity to work in his group and for the guidance received for my research projects. Thank you for all the discussion and for giving me opportunities to present my work in national and international meetings. During the last four years, I have developed myself not only scientifically but also personally.

I would also like to thank Prof. Dr. Matthias Rarey for being my second supervisor and for hosting me at the Center for Bioinformatics.

Thanks to everyone in our group for fruitful discussions and the nice times we have been spending together. Special thanks to Christina de Bruyn Kops for collaborating on the first review we wrote together and the proofreading of some of my manuscripts, and also some work-unrelated talks. Thanks to Nils-Ole Friedrich for his help that I received from him especially during my first year in the group. Thanks to Conrad Stork for the support involving machine learning techniques for the NP-Scout project. Also thanks to Marina Garcia de Lomana, Steffen Hirte and Neann Mathai for their contributions to my different projects. I am very glad that I had chances to collaborate directly with many members of our group. Also thanks to Anke Wilm, Isabel Agea Lorente, Christoph Bauer, and Ningning Fan for all kinds of discussions. Special thanks to Martin Šicho and Méliné Simsir for the time we were in the same office and the encouragement.

Also many thanks to the IT department, Jörn Adomeit and Gerd Embruch for the maintenance of our hardware and software to support my many large calculations.

Thanks to Christina, Neann and Conrad for proofreading this thesis.

I also want to thank Prof. Dr. Gerhard Wolber for reviewing this thesis.

Also many thanks to my parents and sister for their continuous support and understanding.

Last but not least, I want to thank the China Scholarship Council for the PhD scholarship, and the DAAD (German Academic Exchange Service) and the Faculty of Mathematics, Informatics and Natural Sciences (MIN) Graduate School, Universität Hamburg, for their support of my attendance of the 22nd European Symposium on Quantitative Structure-Activity Relationship (EuroQSAR) in 2018.

Contents

Abstract	i
Zusammenfassung	iii
Acknowledgements	v
Contents	vii
1. Introduction	1
1.1. Background	1
1.2. Aims	20
2. Data Resources and Methods	21
2.1. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products	21
2.2. Chemical Data Preprocessing and Molecular Descriptor Calculation	71
2.3. Chemical Space Analysis	71
2.4. Rule-Based Approaches	72
2.5. Machine Learning Methods	73
2.6. Three-Dimensional Shape-Based Similarity Method	75
3. Results	77
3.1. Characterization of Physicochemical and Structural Properties of Natural Products	77
3.2. Machine Learning Method for Assessing Natural Product-Likeness	94
3.3. Scope of 3D Shape-Based Approaches in Predicting the Macromolecular Targets of Structurally Complex Small Molecules	112
4. Concluding Discussion	131
Bibliography	137
Bibliography of this Dissertation's Publications	141
Abbreviations	143
Appendix A	145
Appendix B	173
Appendix C	175
Scientific Contributions	179
Publications	179
Oral Presentations	181
Poster Presentations	181
Awards	182

1. Introduction

1.1. Background

Natural Products (NPs) have a long and successful record of use as components of traditional medicines and herbal remedies. For modern small-molecule drug discovery, NPs remain the most prolific source of inspiration [1]. As presented in a recent statistical analysis of approved drugs from 1981 to 2019, about two-thirds of all small-molecule drugs are related to NPs, including unaltered NPs and NP-derivatives, NPs mimics and/or molecules containing NP pharmacophores [1].

Because of the long history of evolution, NPs have a wide range of bioactivities in different organisms and a large number of NPs are regarded as privileged structures [2,3]. This is related to their high diversity in terms of molecular structures and physicochemical properties. Many NPs have favorable absorption, distribution, metabolism, and excretion (ADME) properties, and some lie outside the general drug-like chemical space [4,5]. Some NPs tend to be highly complex in terms of molecular structure, for example with regard to three-dimensional (3D) molecular shape, stereochemistry and ring systems [6–8].

However, the bottleneck of NP-based drug discovery is the availability of materials for experimental testing. Computational methods have made considerable contributions in different aspects of NP-based drug discovery and provide support throughout different stages of *in silico* early drug discovery [9,10]. These methods have been shown to be able to help researchers to focus on the most promising (plant) materials for experimental testing [11–14].

In this chapter, the importance of NPs as sources of inspiration for drug discovery as well as the role of cheminformatics methods in NP-based research will be highlighted. The scopes and limitations of computational methods in (i) data curation and NP dereplication, (ii) chemical space analysis, visualization, navigation and comparison, (iii) quantification of NP-likeness, (iv) prediction of bioactivity spectra, ADME and safety profiles (toxicity), (v) natural product-inspired *de novo* design and (vi) prediction of natural products prone to cause interference with biological assays, will be discussed in the following timely review (D1).

[D1] Chen, Y.; Kirchmair, J. Cheminformatics in Natural Product-Based Drug Discovery. *Mol. Inf.* **2020**, *39*, 2000171.

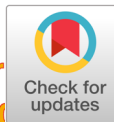
Available at <https://doi.org/10.1002/minf.202000171>.

Y. Chen and J. Kirchmair conceptualized the work. Y. Chen analyzed the literature and wrote the largest part of the manuscript. J. Kirchmair supervised this work.

This article is reprinted from:

Chen, Y.; Kirchmair, J. Cheminformatics in Natural Product-Based Drug Discovery. *Mol. Inf.* **2020**, *39*, 2000171.

This is an open access article published under the Creative Commons Attribution (CC-BY) license.



DOI: 10.1002/minf.202000171

Cheminformatics in Natural Product-Based Drug Discovery

Ya Chen^[a] and Johannes Kirchmair^{*[a, b]}

Abstract: This review seeks to provide a timely survey of the scope and limitations of cheminformatics methods in natural product-based drug discovery. Following an overview of data resources of chemical, biological and structural information on natural products, we discuss, among other aspects, in silico methods for (i) data curation and natural products dereplication, (ii) analysis, visualization, navigation and comparison of the chemical space, (iii) quantification of natural product-likeness, (iv) prediction of the bioactivities

Keywords: cheminformatics · natural products · drug discovery · databases · in silico methods

(virtual screening, target prediction), ADME and safety profiles (toxicity) of natural products, (v) natural products-inspired de novo design and (vi) prediction of natural products prone to cause interference with biological assays. Among the many methods discussed are rule-based, similarity-based, shape-based, pharmacophore-based and network-based approaches, docking and machine learning methods.

1 Introduction

Natural products (NPs) have a long record of use as components of traditional medicines and herbal remedies. Even for modern small-molecule drug discovery they remain the single most prolific source of inspiration.^[1] In fact, about two-thirds of all small-molecule drugs approved between 1981 and 2019 are related, to different extents, to NPs.^[1] Whereas only 5% of the drugs that have been introduced to the market during this timeframe are unaltered NPs, 28% are NP derivatives, and 35% mimic and/or contain a NP pharmacophore.^[1] A highly visible recognition of the relevance of NP-research for public health is the award of the 2015 Nobel Prize in Physiology or Medicine to William C. Campbell, Satoshi Omura, and Youyou Tu for the discovery of two NPs (ivermectin and artemisinin) that led to fundamental improvements in the treatment of diseases caused by parasites.

As a result of evolutionary processes, NPs have a wide range of bioactivities in different organisms. For this reason a substantial number of NPs are recognized as privileged structures.^[2,3] NPs are highly diverse in their molecular structures and physicochemical properties. Many of them have favorable ADME and physicochemical properties; others are clearly beyond what is generally considered as the drug-like chemical space.^[4-6] NPs can be highly complex in terms of molecular structure, in particular with regard to their 3D molecular shape, stereochemistry, ring complexity (macrocycles; bridged or fused ring systems) and conformational space (high number of rotatable bonds; low degree of aromaticity).^[7-9] This poses fundamental challenges to 3D cheminformatics methods for which reasons the development of force fields and algorithms for the prediction of the protein-bound conformations of such complex molecules remains one of the most actively pursued research topics in cheminformatics.^[10-15]

The real bottleneck of NP-based drug discovery, however, is the availability of materials for testing. The sourcing process can be complex, lengthy and costly, and transport across borders may prove legally challenging.^[16] Once the material has arrived at its destination, the production of extracts, the in vitro testing for bioactivity, the identification

and isolation of the bioactive compounds from these complex mixtures, the determination of the mode of action, the resupply of compounds of interest (e.g. through partial or total chemical synthesis), and the profiling of their pharmacological, pharmacokinetic and toxicological properties all require expertise, substantial efforts, time and funds, and there is no guarantee of success.^[4,16,17]

Computational methods can make substantial contributions to NP-based drug discovery and support experimentalists throughout the hit discovery, hit-to-lead and lead optimization phases.^[18,19] They have been shown to be particularly powerful, not just in identifying bioactive NPs, but also in prioritizing (plant) materials for testing,^[20-23] hence helping experimentalists to focus their resources on the most promising materials. Computational methods are also employed, for example, in (i) data curation and NP dereplication, (ii) chemical space analysis, visualization, navigation and comparison, (iii) quantification of natural product-likeness, (iv) prediction of bioactivity spectra, ADME and safety profiles (toxicity), (iv) natural products-inspired de novo design and (v) prediction of natural products prone to cause interference with biological assays.


Compared to the costs involved in experimental approaches, the funds required for in silico experiments


[a] Y. Chen, J. Kirchmair
Center for Bioinformatics (ZBH), Department of Computer Science,
Faculty of Mathematics, Informatics and Natural Sciences, Uni-
versität Hamburg, 20146 Hamburg, Germany
Tel.: +43 1-4277-55104

E-mail: johannes.kirchmair@univie.ac.at

[b] J. Kirchmair
Department of Pharmaceutical Chemistry, Faculty of Life Sciences,
University of Vienna, 1090 Vienna, Austria
Tel.: +43 1-4277-55104

E-mail: johannes.kirchmair@univie.ac.at

 Special Issue "7th Strasbourg Summer School in Chemoinformatics"
(Dragos Horvath)

 © 2020 The Authors. Published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

seem almost negligible. An in-house high-performance computing facility is no longer essential. Today, calculations can be run (if at all needed) at very large scales in the cloud, at moderate cost and low complexity. Merely software license fees remain a substantial cost factor and have constantly increased throughout recent years. At the same time, we are now seeing a growing number of powerful open-source tools becoming available, much like what has been quite common to the field of bioinformatics. Some of the most outstanding software in this context are RDKit^[24] and CDK^[25,26] (both are open-source toolkits for cheminformatics), KNIME^[27] (an open-source analytics platform), and scikit-learn^[28,29] (an open-source Python module for machine learning).

With this review, we aim to provide a succinct but comprehensive overview of the scope and limitations of cheminformatics methods in NP-based drug discovery in a format that is accessible to researchers from different domains with an interest in drug discovery. The discussion covers a large number of state-of-the-art methods in cheminformatics as well as data resources relevant to NP-based drug discovery.

2 Natural Products Collections Relevant to Computer-guided Natural Products Research

2.1 Virtual Natural Products Collections

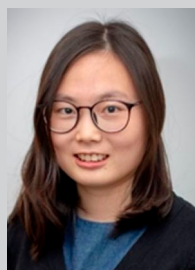
The last decade has seen a steep increase in databases providing access to chemical, biological, pharmacological, toxicological and structural data on NPs. We recently conducted comprehensive surveys of databases that are particularly relevant to NP-based drug discovery.^[6,30,31] As a minimum requirement, any of the more than 30 databases surveyed feature a chemistry-aware web interface for searching and browsing molecular structures. Most of the databases also offer free bulk download, enabling virtual screening and other applications. From these studies we

gathered that the total number of NPs for which their structures can be obtained via bulk download from free databases is in excess of 250k, approaching 300k.

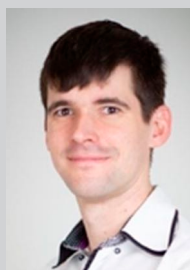
Unfortunately, the half-life of many (NP) databases is short; only few of them are sustainably managed and under continued development. Data quality is always of concern, but when it comes to NPs, extra caution should be exercised, in particular when using the data with computational methods relying on the accurate representation of 3D molecular structures. This is because stereochemical information on NPs is fairly commonly inaccurate or incomplete.

Virtual NP databases can be categorized into (i) encyclopedic and general NP databases, (ii) databases enriched with NPs used in traditional medicines, (iii) specialized databases focused on specific habitats, geographical regions, organisms, biological activities, or even specific NP classes. The largest of all free NP databases is Super Natural II,^[32] which consists of more than 325k NPs. The database can be queried via a chemistry-aware web interface but bulk download is not officially supported. Among the most outstanding free, downloadable resources is the Universal Natural Products Database (UNPD),^[5] which lists more than 200k NPs from all forms of life. Unfortunately, this database appears to no longer be hosted. Further large databases include the TCM database@Taiwan,^[33] which lists more than 60k NPs found in Chinese medical herbs, the Natural Product Atlas,^[34,35] offering data on over 25k NPs from bacteria and fungi, and the Collective Molecular Activities of Useful Plants (CMAUP) database,^[36] a collection of over 47k NPs from more than 5600 plants with their biological activities information.

In contrast to information on molecular structures, data on the biological activities and protein-bound conformations of NPs remain sparse. By overlapping our set of approximately 250k NPs with the full ChEMBL database (a database providing bioactivity data on approximately 2 Million compounds),^[37,38] we found that only about 16% were present in the ChEMBL database and had at least one



Ya Chen is a Ph.D. student with Ass.-Prof. Johannes Kirchmair at the Center for Bioinformatics (ZBH) of the Universität 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.



Johannes Kirchmair is an assistant professor in cheminformatics at the Department of Pharmaceutical Chemistry of the University of Vienna and head of the Computational Drug Discovery and Design Group (COMP3D). He also is a group leader at the Center for Bioinformatics (ZBH) of the Universität Hamburg. After earning his PhD from the University of Innsbruck (2007), Johannes worked in different capacities at Inte:Ligand GmbH (Vienna), BASF SE (Ludwigshafen), the University of Cambridge and ETH Zurich. He also held a junior professorship in applied bioinformatics at the Universität Hamburg (2014 to 2018) and an associate professorship in bioinformatics at the University of Bergen (2018 to 2019).

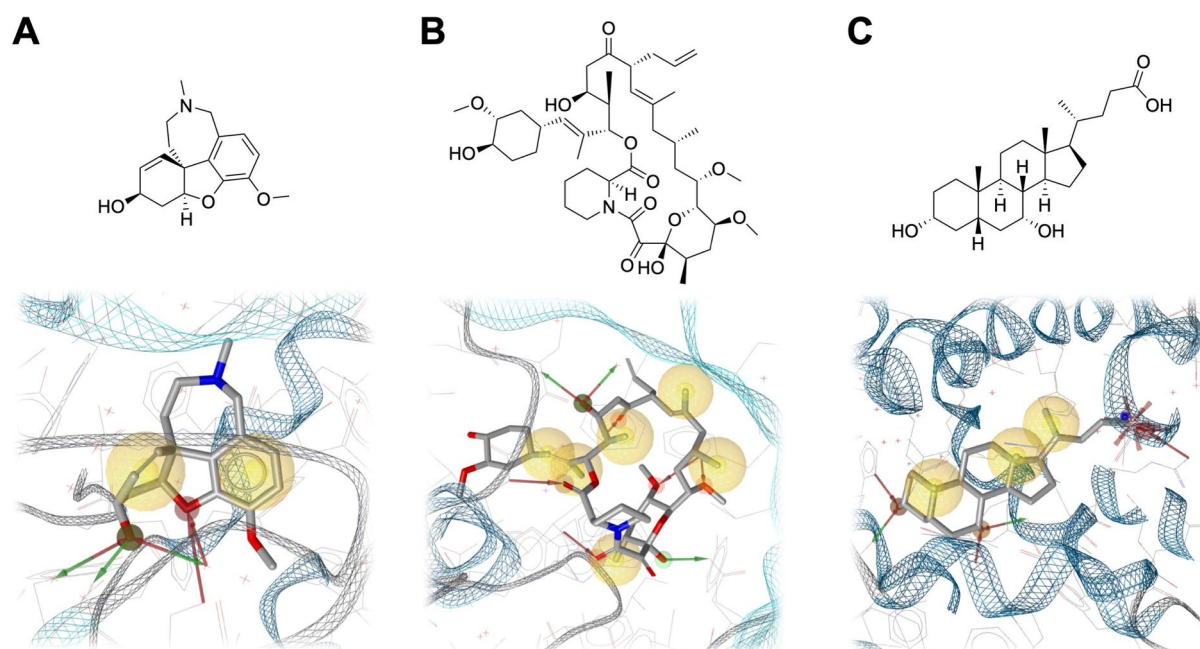


Figure 1. Examples of approved NP drugs and how they bind to their target proteins: (A) (-)-galantamine, an acetylcholinesterase inhibitor approved for the treatment of Alzheimer's disease (PDB ID 1DX6), (B) tacrolimus, a macrocyclic immunosuppressant targeting the immunophilin FKBP-12 (FK506 binding protein; PDB ID 1FKF) and (C) chenodeoxycholic acid, an endogenous bile acid that is used for the treatment of hypocholesterolemia. Chenodeoxycholic acid stimulates the farnesoid X receptor (FXR; PDB ID 6HL1). Carbon atoms grey; oxygen atoms red; nitrogen atoms blue. Hydrogen bonds formed between the ligand and the protein or water molecules are visualized by red arrows (acceptors on the ligand side) and green arrows (donors on the ligand side); hydrophobic features are visualized as yellow spheres, and negative ionizable features as red stars. Visualization and pharmacophore perception with LigandScout.^[39]

bioactivity annotation.^[31] Likewise, by overlapping the NP dataset with all small-molecule ligands represented in the Protein Data Bank (PDB), we found that for only about 2000 NPs at least one co-crystallized X-ray structure of high quality is available.^[6] The X-ray structures of three NPs approved as drugs and bound to their target proteins are shown in Figure 1.

Since the publication of our recent works,^[30,31] more than one dozen new NP databases have appeared and existing ones have been updated. However, only few of these databases offer bulk download of molecular structures. Among the most relevant databases to mention is the Marine Natural Library,^[40] which allows the download of the full dataset of more than 14k marine NPs. In early 2020, a new database was introduced which its authors claim to be the world's largest collection of NPs.^[41] It should be noted that this database combines data from resources of which some are known to also include substantial numbers of NP derivatives and analogs, and that the data will require additional curation for most applications in cheminformatics.^[41]

The reader is referred to refs. [30,31,41–45] for additional information on NP databases relevant to cheminformatics.

2.2 Physical Natural Products Collections

Today, most of the hundreds of compound suppliers worldwide provide comprehensive information on the molecular structures (and other properties) of their compounds for the purpose of virtual screening and other applications free of charge. The majority of the commercial compound collections are dominated by synthetic compounds. By overlapping a comprehensive collection of more than 250k NPs (which we compiled by curating and merging all of the NP datasets available to us^[31]) with the 7.3 million in-stock compounds listed in the ZINC database^[46,47] (a comprehensive database of compounds that are available from various commercial sources and research institutes), we found that only about 10% of the known NPs (approximately 25k) are readily obtainable for experimental testing.^[31] This confirms that the availability of materials for experimental evaluation represents the bottleneck in NP-based drug discovery. Note that by allowing minor structural deviations between NPs and purchasable compounds, meaning the inclusion of mainly NP derivatives and analogs, the number of readily obtainable compounds increases by roughly 10k to 30k.^[31] It is also worthwhile mentioning that the majority of the readily obtainable NPs have physicochemical properties that are considered favorable in the context of drug discovery. In fact, more than half

of them are fragment-sized (molecular weight below 300 Da),^[31] hence offering ample opportunities for optimization.

Purified NPs are available from more than 100 commercial providers worldwide^[31] but only a dozen of these companies offer more than 5000 NPs. Pure collections of genuine NPs are rare whereas mixed catalogues are commonplace. In these mixed catalogues, however, genuine NPs, NP derivatives and NP analogs are rarely labeled as such. Surprisingly often there is no mention of NPs found on the websites of compound providers, even of those vendors that offer substantial numbers of different NPs. Therefore, tools for identifying NPs and NP-like compounds can be of high value to NP-based drug discovery (see Section 6 for details).

The discussion of catalogue sizes should not obscure the importance of compound diversity with respect to physicochemical, structural and biological properties. In this context it is encouraging to know that the (above-mentioned) 25k readily purchasable NPs cover more than 5700 Murcko scaffolds. We also found that the readily purchasable NPs give a good representation of all of the major NP classes, such as alkaloids, steroids and flavonoids.^[6]

3 Computational Methods for Structure Elucidation and Dereplication of Natural Products

The sourcing of materials for the extraction and isolation of NPs are expensive and time-consuming, and with increasing knowledge of NPs, the chances for finding novel compounds are diminishing. In order to enable the efficient use of the available experimental resources, analytical and computational methods are utilized in tandem in order to identify known NPs as well as NPs with undesirable properties at the earliest possible point in time.^[44] An important component in this interplay of technologies are databases providing measured analytical data (e.g. bioactivities, chromatographic data, mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy data) for known NPs and their interrogation with computational methods. However, even the largest of these databases cover only a small fraction of the known NPs, for which reason computational methods are increasingly being employed also for the prediction of MS fragmentation and NMR spectra, sometimes in combination with structure generators.^[44]

There are elaborate algorithms in place which allow the transformation of spectral data into representations (reduced to peak lists, numerical vectors, trees or others) that enable the efficient comparison of spectra and ranking according to their similarity. In other words, these methods have the capacity to identify spectra derived not only from the same compounds but also from structurally related compounds. This means that the applicability of these

methods goes beyond known NPs and that they can provide, for example, valuable hints on chemical classes and functional groups. However, such analyses still require manual interaction by an expert, hence limiting automation.^[48]

A main approach to computer-assisted dereplication is the combination of analytical data with multivariate data analysis.^[44] Using dimensionality reduction techniques such as principal component analysis (PCA), clustering methods, and/or discrimination analysis can help to identify interesting NPs in complex mixtures, e.g. NPs in extracts that are unique to a particular organism of interest.^[49,50]

Systems for computer-assisted structure elucidation (CASE) aim to identify the correct structure of a compound of interest based on the available spectroscopic data.^[51] More specifically, CASE systems enumerate the structures that are consistent with the experimental (spectroscopic) data and rank them according to their probability. Ideally, CASE systems work in a fully automated fashion, at low error rates. Elaborate CASE systems also take stereospecific NMR data and/or calculations based on density functional theory into account and hence can be used for the assignment of stereochemical properties to NP structures.^[51]

Machine learning approaches enjoy high interest in NP dereplication. For example, in a recent study the capacity of machine learning algorithms to assign NPs to eight NP classes (such as chromans) based on ¹³C NMR spectroscopy data was explored.^[52] The best performance was obtained with an XGBoost classifier. For most NP classes, more than 80% of the compounds of a test set were correctly assigned. Another study successfully employed a convolutional neural network-based approach for the rapid identification of new NPs from a filamentous marine cyanobacterium.^[53]

A different approach is taken by the NP-StructurePredictor.^[54] Based solely on targeted molecular weights derived from *m/z* values obtained by liquid chromatography-MS, this tool produces a rank-ordered list of likely NP structures. In order to do so, the tool features a structure generator that can combine the different scaffolds and decorations (which draws from a large NP database), and that can infer structures from structurally related scaffolds.

For more information on experimental and computational methods for NP dereplication readers are referred to recent reviews on this topic, for example, refs. [44,48,55,56].

4 Computational Analysis of the Physicochemical and Structural Properties of Natural Products

Cheminformatics has been playing a key role in the characterization of NPs by their physicochemical and structural properties, and in the comparison of NPs with

small-molecule drugs, drug-like compounds and other types of (organic) molecules. NPs cover a much broader chemical space than synthetic compounds and they populate also areas in chemical space that are generally not (or only with great difficulties) synthetically accessible.^[6,8,19,57,58] The structural uniqueness (and complexity) of some NPs could allow them to target macromolecules that are otherwise undruggable.^[16]

NPs are on average heavier and more hydrophobic than synthetic drugs and synthetic, drug-like compounds.^[59] Their structural complexity is also often higher, in particular with regard to stereochemistry (commonly quantified by the number of chiral centers,^[57,59–66] the number of fraction of Csp³ atoms,^[6,8] and/or the number of bridgehead atoms in ring systems^[67]) and 3D molecular shape.^[8,68]

NPs show an enormous diversity of ring systems, in particular of aliphatic systems.^[68,57,63,65] One study showed that 83% of core ring scaffolds of NPs are absent in commercially available screening databases.^[69] With regard to atom composition, two of the most discriminative features of NPs over synthetic compounds are the (on average) low number of nitrogen atoms and high number of oxygen atoms.^[57,59,62–64] Nevertheless, a clear majority of the known NPs, and even more so in physical NP libraries, are drug-like.^[6]

NPs from different kingdoms have distinct physicochemical and structural properties.^[66,70–76] For example, NPs with macrocycles or long aliphatic chains are more commonly to marine species than terrestrial species.^[74] Also bacteria produce many macrocyclic NPs.^[75] Their NPs are characterized by a high proportion of heteroatoms and, related to this, a high diversity of functional groups.^[76]

5 Computational Methods for the Assessment of the Structural Diversity of Natural Products

NPs are unrivalled in terms of structural diversity, a fact which is also reflected on a fragment level.^[77] Most of the studies assessing the structural diversity of NPs and comparing them to that of synthetic compounds make use of the concept of molecular frameworks (scaffolds) introduced by Bemis and Murcko.^[78] In recent work, Ertl and Schuhmann^[75] show an intuitive visualization of scaffolds characteristic to NPs and compare them with those of synthetic compounds. They also provide a comparison of scaffolds frequently observed in NPs produced by bacteria, plants, fungi or animals. Rule-based methods offer a different angle towards NP diversity analysis. They allow, for example, the automated assignment and assessment of the major NP classes.^[6]

A powerful tool for the intuitive, visual analysis of the structural diversity of sets of compounds is Scaffold Hunter.^[79,80] The Java-based, open source software features a graphical user interface and multiple clustering algorithms. Scaffold Hunter is based on the idea of the

hierarchical representation and classification of molecular scaffolds (“scaffold tree”). An early version of this tool formed the basis of the structural classification of NPs (SCONP), a method for charting the chemical space of NPs.^[81]

One of the most commonly employed techniques for mapping the chemical space is PCA,^[6,58,59,64,73,82,83] which projects high-dimensional data into a low-dimensional space for improved interpretability, while keeping information loss to a minimum. The most relevant result of PCA and starting point for interpretation is the PCA scatter plot, which shows the distribution of the data points in the low-dimensional space. When interpreting a PCA scatter plot it is very important to understand and consider the proportion of variance explained by the shown (two or three) principal components. Only if the proportion of variance explained is sufficiently high, the observed distribution of the data points is informative. This is typically not the case for PCAs based on molecular fingerprints; physicochemical property descriptors usually give better results with PCA.

To avoid the need for the recalculation of the principal components as new compounds are added to the datasets, a method named ChemGPS^[84] was developed and extended for use with NPs (“ChemGPS-NP”^[85]). The method utilizes predefined rules in combination with selected molecular structures to render a “global drugspace map” into which new structures are projected based on predicted PCA scores. ChemGPS-NP has been used in several studies for mapping the chemical space of small molecules,^[71,86] for mode of action prediction,^[87] and for the analysis of structure-activity relationships.^[86,88]

Also self-organizing maps and generative topographic maps have been regularly utilized for comparing the molecular structures of NPs with those of drugs, and for visualizing the structural diversity of fragment-sized and non-fragment sized NPs.^[66,89,90] One interesting observation from these analyses is a high degree of resemblance of NPs and synthetic drugs in term of their pharmacophore features, despite profound differences in chemical structure.^[90]

Further powerful methods for dimensionality reduction include T-distributed Stochastic Neighbor Embedding (t-SNE)^[91] and the recently introduced Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) method.^[92] t-SNE produces plots where, overall, similar objects are located in close proximity and dissimilar objects are modeled by distant points. t-SNE can produce visualizations that are superior to those from PCA but the method does not scale well with the size of data sets. UMAP is conceptually related to t-SNE and produces similar results but it is faster.

The research group of Medina-Franco has been developing several methods for the intuitive characterization, visualization and comparison of compound collections, with focus on NP databases. For example, they developed the Consensus Diversity Plot (CDP),^[93] which allows the compar-

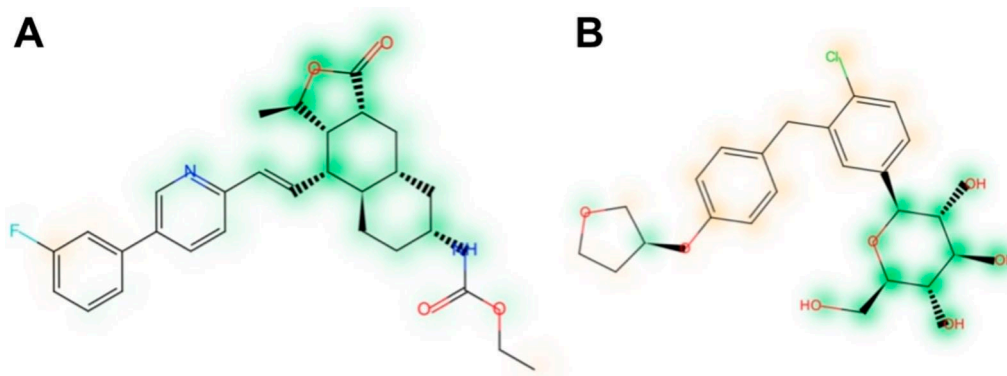


Figure 2. Similarity maps of (A) vorapaxar and (B) empagliflozin. Green-highlighted atoms contribute to the classification of a molecule as a natural product; orange-highlighted atoms contribute to the classification of a molecule as a synthetic compound. Adapted from [59] (CC BY 4.0; <https://creativecommons.org/licenses/by/4.0>).

ison of datasets by a single, straightforward 2D plot representing the median (or other) values of four key properties of choice (e.g. physicochemical property, molecular diversity, scaffold diversity). Each dataset is represented by a single data point. The data point is positioned in the 2D plot according to two properties of choice represented by the x and y axes. The third property of choice is represented by color coding of the data points, and the fourth one (intuitively, this would be the database size) is represented by the size of the data point. The method has been used for the visual comparison of multiple small-molecule databases^[83,94–96] and is accessible via a web service.^[93]

Recently, researchers from the same group reported the development of a new method for the representation of the chemical space of compound databases by a single fingerprint called Statistical-Based Database Fingerprint (SB-DFP).^[97] The SB-DFP is widely applicable and can be derived, in principle, from any molecular fingerprint and for any reference set. The SB-DFP is generated by comparing the binomial distributions of features of the molecular fingerprint of choice among the compounds of a dataset of interest and that of a reference dataset. Only bits for which significantly higher “on” rates are observed in the molecular fingerprint among the compounds in the dataset of interest (than in the reference set) will be set to “1” in the SB-DFP. The SB-DFP was utilized for assessing and visualizing the similarity of the chemical space of sets of NPs and synthetic compounds, confirming that NP collections cover ample chemical space that remains to be explored (more thoroughly) in the context of drug discovery.

6 Computational Methods for the Assessment of Natural Product-likeness

Computational tools are able to discriminate NPs and NP-like compounds from synthetic compounds with high

accuracy, and they are also able to quantify the NP-likeness of compounds. As such they are commonly applied to compound design, library design, the selection of NPs (and NP derivatives and analogs) from mixed compound collections, and for compound prioritization.^[59,98]

One of the most established approaches is the NP-Likeness Score developed by Ertl et al.^[99] Employing Bayesian statistics, this score quantifies the NP-likeness of compounds based on the similarity of their fragments with those of known NPs. The NP-Likeness Score has been re-implemented in different software and platforms, with some modifications.^[100–103] Further approaches include a conceptually related method employing extended connectivity fingerprints (ECFPs)^[98] as well as a rule-based approach.^[104] More recently, we developed NP-Scout,^[59] a tool for identifying NPs and NP-like compounds in large sets of molecules. The random forest classifiers are trained on a large collection of known NPs and synthetic compounds. On a representative test set, a classifier based on MACCS keys obtained an area under the receiver operating characteristic curve (AUC) of 0.997 and a Matthews correlation coefficient (MCC) of 0.960. NP-Scout makes use of similarity maps, which highlight areas in a molecule that contribute to the prediction of a molecule as NP or synthetic compound (Figure 2). NP-Scout is accessible via a free web service.^[105]

Most recently, the Natural Compound Molecular Fingerprint (NC-MFP) was introduced as a new approach of describing in particular the structural features of NPs in terms of the scaffolds and fragments they are composed of.^[106] The NC-MFP was shown to outperform established fingerprints in discriminating NPs from synthetic compounds.

7 Computational Methods for the Identification of Bioactive Natural Products

Computational methods have a strong track record in the identification of bioactive NPs. The entire range of virtual screening methods has been applied for NP research, from simple, fast methods based on 2D molecular fingerprint similarity to more complex, 3D methods based on molecular shape similarity, pharmacophore models, molecular interaction fields, or docking. More recently, machine learning approaches have become a mainstay in virtual screening for bioactive NPs.^[107]

In particular 3D virtual screening methods are challenged by the structural properties of many NPs such as high degrees of conformational flexibility, the complexity of their molecular shapes and ring systems (notably macrocycles), insufficiencies of molecular force fields primarily parameterized for synthetic compounds, and uncertainties related to protonation states, tautomerism and oxidation states (for example, the possible involvement of polyphenols in redox cycles is often disregarded). One approach to reduce the structural complexity of NPs is to remove the sugars and sugar-like components from NPs in cases where they are deemed not to be essential for bioactivity.^[66,108] This can be done, for example, by use of defined (SMARTS) patterns.^[6,100]

Given the sparsity of available structural data, docking of NPs to the structures of macromolecules can pose a profound challenge. This is because docking algorithms and scoring functions are highly sensitive even to very small changes in 3D structure such as those commonly induced by ligand binding (including solvent effects). However, also this hurdle may be overcome by the prudent use of homology modeling techniques, induced fit docking approaches, and/or molecular dynamics simulations. In the case of highly flexible proteins, docking against multiple, representative protein structures ("ensemble docking") may be a good way forward (not only for virtual screening but also for binding mode prediction).^[109,110] Diligence and patience will certainly be required and, above all, checks of the plausibility of a hypothesis using all available information can help to piece the puzzle together.

More often than in virtual screening-docking algorithms produce good results in binding mode prediction.^[111] Provided that the NP of interest is not excessively large or flexible (as a rough guide, not exceeding 35 heavy atoms or eight rotatable bonds), that the ligand binding site is well-defined (i.e. not overly shallow, not solvent-exposed), and that the interaction between the binding partners involves two or more directed interactions, there is a good chance that a sufficiently accurate binding pose can be obtained that offers crucial insights for the development of optimization strategies. Binding pose prediction is more feasible than virtual screening because it allows to largely disregard the most challenging aspect of docking, which is the scoring of compounds according to their binding affinity,

and it allows researchers to focus their effort on one specific ligand-target pair. Importantly, in particular in the context of NP research, docking enables the rationalization of stereoselectivity in ligand binding (and other processes, such as metabolism). The importance of using the correct stereochemical information with 3D approaches, especially with docking, cannot be overstated.

In the following paragraphs we briefly discuss representative examples of studies in which virtual screening was successfully employed for the identification of bioactive NPs. For more comprehensive discussion of applications, the reader is referred to excellent reviews.^[18,112]

Using katsumadain A (a diarylheptanoid inhibiting influenza neuraminidase) as a template for 3D molecular shape-based screening, a number of structurally distinct NPs were identified that inhibit the viral enzyme with IC_{50} values in the submicromolar to low micromolar range (for example artocarpin (1), which is depicted in Figure 3).^[113] In another study, pharmacophore-based virtual screening was combined with a shape-based approach in order to identify activators of the G protein-coupled bile acid receptor 1 (GPBAR1).^[114] In addition to several NP databases also a collection of synthetic compounds was screened. Among the 14 selected NPs eight (57%) obtained a measured receptor activation of at least 15% at 20 μ M concentration. Two of these compounds, farnesiferol B (2) and microlobidene (3), are based on molecular scaffolds that had not yet been associated with GPBAR1 modulation. Both compounds were reported to have EC_{50} values of approximately 14 μ M. Among the 19 selected synthetic compounds, only two were active (applying the identical activity threshold).

Influenza neuraminidase has also been successfully addressed by docking. For example, a database of NPs related to plants endogenous to Malaysia was screened for potential inhibitors of influenza neuraminidase.^[20] From the five plants with the highest hit rates in docking, twelve NPs with moderate inhibitory activity on influenza neuraminidase were identified by experimental testing (one example is rubraxanthone (4)), four of which had been ranked by docking among the top-100 compounds in the hit list.

A pharmacophore approach was utilized to screen a collection of 10k NPs related to traditional Chinese medicine for compounds targeting the farnesoid X receptor (FXR), a transcription factor involved in inflammatory liver diseases.^[115] Screening results indicated a high likelihood of activity of lanostane triterpenes from the mushroom *Ganoderma lucidum*. Several of these lanostanes were isolated and subjected to experimental testing in a reporter gene assay. Five lanostanes showed a dose-dependent induction of FXR with EC_{50} values in the low micromolar range, the most active ones being ergosterol peroxide (5) and ganodermanontriol (6).^[21]

Rupp et al.^[116] explored a number of different machine learning approaches in order to identify NP derivatives that selectively activate the peroxisome proliferator-activated receptors (PPAR γ). The authors focused on the use of

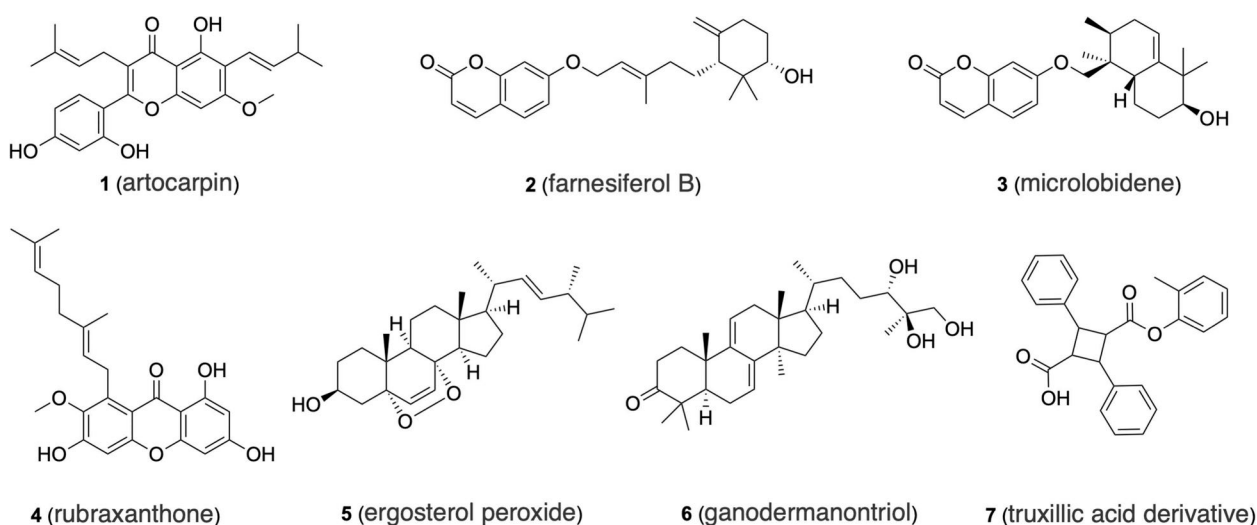


Figure 3. Examples of natural products and natural product derivatives identified by virtual screening.

Gaussian process models (with different kernels) that they employed to learn pharmacophoric patterns from a medium-sized set of synthetic PPAR γ ligands. By screening and ranking several hundred thousand commercially available compounds, the authors identified a truxillic acid derivative (7) as a selective activator of PPAR γ (EC_{50} = 10 μ m).

Another study from the same lab^[117] employed machine learning-based virtual screening for the identification of mimetics of the Alzheimer drug (–)-galantamine (Figure 1). Like for many Alzheimer drugs, the therapeutic efficacy of (–)-galantamine is linked to activities on multiple proteins rather than a single one. In the search for efficacious compounds it is hence important to consider polypharmacology. To this end, Grisoni et al. employed the machine learning-based target prediction models SPIDER and TIGER (which are discussed in more detail in the next section) to identify (in this case synthetic) compounds with bioactivity spectra that are comparable to that of (–)-galantamine. Using these models, they selected 20 compounds from a set of more than 3 Million purchasable compounds for testing. Among the selected compounds, several showed interesting activities *in vitro*. Two compounds of small size were shown to have polypharmacological profiles that are considered to be favorable for the treatment of Alzheimer's disease.

8 Computational Methods for the Prediction of the Macromolecular Targets of Natural Products

Knowing the macromolecular target(s) of small molecules is of utmost importance to the assessment of the pharmacological efficacy and safety of compounds, and for their further development. However, even for a substantial

number of marketed drugs the mode of action is unknown or only vaguely understood. The road to the experimental identification of the target(s) of small molecules can be very lengthy and expensive, and there is a good chance to be met by disappointment on the way, for example, when it becomes clear that “the target” of a supposedly innovative compound is an established drug target or, worse, a protein known to be not a viable drug target. Computational approaches are hoped to make a significant contribution to making mode of action identification more efficient and there is an increasing body of evidence that some of these hopes are becoming reality (as will be discussed below).

In silico target prediction can be regarded as a large-scale application of virtual screening (see the previously discussed study of Grisoni et al.^[117]), in the way that one, several or many compounds are screened against the widest possible set of macromolecules. A plethora of methods and models have been reported in recent years^[118–121] and they have become established as important tools in early drug discovery. Related to the challenges involved in docking and structure-based methods in general (in particular, the limited coverage of macromolecules by the available structural data), most approaches for target prediction are ligand-based.

Ligand-based methods cover the full range from straightforward similarity-based approaches to complex machine learning and network-based approaches. Surprisingly, despite today's abundance of computational methods for target prediction, our understanding of the value of these methods under real-world conditions remains limited.^[122] This is primarily because of the (in general) prohibitive costs involved in the experimental, systematic, prospective evaluation of such models, but also because of the partly insufficient, superficial retrospective validation protocols that are regularly employed.^[122,123] To our best knowledge, the only computational method for which a

systematic experimental validation has been reported so far remains the well-known Similarity Ensemble Approach (SEA).^[124–126] One may rightly argue that validating models on existing data generally leads to an overestimation of how well a model will perform under real-world conditions, however, there is at least one more important point to consider when judging the value of target prediction approaches based on retrospective validation studies: under real-world conditions, researchers will rarely face the situation where no hints on a compound's target are available at all. A scenario where a substantial amount of information is available on a compound of interest, e.g. phenotypic assay readouts with different cell lines or data for structurally related compounds, is more likely. By adding up all of the available information it is likely that many false-positive predictions can be ruled out, hence leaving much fewer candidate targets to be investigated experimentally.

In a recent, in-depth study of the performance and scope of a similarity-based approach and a machine learning approach for predicting the targets of small molecules, we show that the reliability of predictions of either approach strongly depends on the structural relationship between the compounds of interest and compounds represented in the training set (or knowledge base).^[123] This fact needs to be carefully considered when working with NPs, given the fact that models for target prediction are mostly designed for, and trained on, measured data for synthetic compounds.

In the same study we found that, surprisingly, with the currently available data, the similarity-based approach generally outperformed the machine learning approach. While a direct comparison of these two approaches should, for several reasons, be considered with great caution, the results suggest that the simple similarity-based approach is a good choice, in particular also when taking into account model interpretability. This is also reflected by the good performance of other established, similarity-based models such as SwissTargetPrediction.^[127]

Most NPs are structurally distinct from more conventional, synthetic compounds, which account for the bulk of the measured activity data. More complex similarity-based methods that compare molecules based on their 3D molecular shape are designed to recognize such distant structural similarity but until recently it was unclear how well these methods would work in practice. We systematically explored the capacity of ROCS,^[128,129] a leading, shape-based screening engine that also takes into account chemical feature distributions, to identify the macromolecular targets of "complex" small molecules based on a knowledge base of "non-complex" compounds with measured bioactivity data.^[130] For the purpose of this work, we defined molecules as "complex" if they are either (very) large in size (45 to 55 heavy atoms) or macrocyclic (and large). In contrast, we defined molecules as "non-complex" if they were small in size (15 to 30 heavy atoms). A total of

28 pharmaceutically relevant targets were studied. For each of the targets a diverse set of 10 complex small molecules was automatically generated. A single, low-energy conformation of each of these molecules was used as a query for screening with ROCS against a multi-conformational knowledge base. The knowledge base represents 3642 targets with a total of 272 640 non-complex small molecules. This study found that ROCS correctly ranked at least one known target among the top 10 positions (out of a list of 3642) for up to 37% of the 280 complex small molecules serving as queries. Considering the dissimilarity of the queries and the compounds in the knowledge base, this performance is remarkable. It indicates that target prediction is possible for a substantial number of challenging complex molecules. Note that researchers will be able, in many cases, to strongly reduce the number of target candidates based on expert knowledge and available information. Among the 280 complex small molecules were at least 31 known, complex NPs and NP-like compounds. For these compounds, the top-10 success rate was lower (23% vs. 37%). This is related to the fact that the median Tanimoto coefficient based on Morgan2 fingerprints of the complex NP (or NP-like compound) and the closest non-complex small molecule in the knowledge base is only 0.13. For pairs of compounds sharing such a low degree of similarity it can be expected that their binding modes are distinct, which is generally beyond the scope of ligand-based methods. In summary, taking into account capacity of these methods and their low demand in computational power, we believe it is worthwhile using these methods in any case as valuable ideas may emerge from their use.

Besides 3D similarity-based approaches, also 3D pharmacophore-based approaches are regularly used for target prediction in the context of NP research. One example is a profiling study in which secondary metabolites isolated from the medical plant *Ruta graveolens* were screened against a battery of more than 2000 pharmacophore models representing over 280 targets.^[131] From this in silico screen, among other bioactive NPs and interactions, arborinine was identified as an inhibitor of acetylcholinesterase (measured $IC_{50} = 35 \mu M$).

In recent years the models for NP target prediction which have seen most interest certainly are those based on machine learning. Notable examples include SPiDER,^[132] TIGER,^[133] and STarFish.^[134] SPiDER uses self-organizing maps in combination with "fuzzy" molecular descriptors that allow for extending its usage to NPs.^[135,136] The model was instrumental in the identification of 5-lipoxygenase, PPAR γ , glucocorticoid receptor, prostaglandin E2 synthase 1, and FXR as targets of the macrolide archazolid A,^[137] and it correctly predicted prostanoid receptor 3 as a target of dolicolide, a 16-membered depsipeptide.^[138] SPiDER also successfully identified the targets of several fragment-like NPs such as (i) sparteine, for which the kappa opioid receptor, p38 α mitogen-activated protein kinase, muscarinic and nicotinic receptors were experimentally confirmed

as targets,^[3] (ii) DL-goitrin, for which the pregnane X receptor and the muscarinic M1 receptor were experimentally confirmed as targets,^[139] (iii) isomacrin, for which the platelet-derived growth factor receptor and the adenosine A₃ receptor were experimentally confirmed as targets,^[139] and (iv) graveolinine, for which cyclooxygenase-2 and the serotonin 5-HT_{2B} receptor were experimentally confirmed as targets.^[139]

Building on predictions from SPiDER, the Drug-Target Relationship Predictor (DEcRyPT)^[140] employs random forest regression in order to generate a refined list of likely macromolecular targets. Use of DEcRyPT led to the successful identification of 5-lipoxygenase as a target of the ortho-naphthoquinone β -lapachone.^[140] The hydroquinone form of β -lapachone was confirmed as a nanomolar inhibitor of 5-lipoxygenase.

TIGER is conceptually related to SPiDER. However, it employs modified CATS descriptors and uses a different method for scoring the predicted targets (taking into account ensemble similarity). TIGER successfully identified the orexin receptor, glucocorticoid receptor, and cholecystokinin receptor as targets of the marine NP (\pm)-marinopyrrole A.^[133] The model also rightly predicted, among other proteins, estrogen receptors α and β as targets of the stilbenoid resveratrol.^[141]

STarFish is a stacked ensemble approach for target prediction trained on synthetic compounds. Various machine learning algorithms were explored as part of the development process. The best stacking approach identified by the authors used molecular fingerprints as input for a random forest model and a k-nearest neighbors model (level 0). The probabilities predicted by these two models for each of the targets are then used as input for a meta-classifier based on logistic regression (level 1). The stacking approach was found to perform substantially better on a test set of NPs (ROC AUC 0.94; BEDROC score 0.73) than the individual models (AUCs between 0.70 to 0.85; BEDROC scores between 0.43 and 0.59).^[134]

Also network approaches focused on the prediction of the macromolecular targets of NPs have been reported. For example, Cheng and co-workers developed statistical network models in order to link NPs to anti-cancer targets^[142] and proteins involved in aging-associated disorders.^[143]

Most recently, multi-task deep neural networks were trained on medical indication data and employed for identifying privileged molecular scaffolds in NPs (in this case, scaffolds for which multiple NPs built on the identical scaffold are active in the same indication).^[144] Based on the predictions of these models, a privileged scaffold dataset for 100 indications was compiled that could serve as a starting point for NP-based drug discovery.

For additional information on this topic, the reader is referred to refs. [18,19,145].

9 Computational Identification of Natural Products Likely to Interfere with Biological Assays

The inclination of NPs to cause interference with biological assays continues to pose a significant challenge to the experimental screening of NPs.^[146,147] The flavonoid quercetin, a known aggregator and pan-assay interference compound, gives an illustrative example of the scale of the problem: as of July 28, 2020, the PubChem Bioassay database listed quercetin as conclusively active in more than 800 unique bioassays, which represents a hit rate of more than 50% (among all conclusive assay outcomes).

By far the most commonly observed mechanism of assay interference is aggregate formation, which occurs under specific assay conditions.^[148] Further relevant mechanisms are covalent binding, redox-cycling, membrane disruption, metal chelation, interference with assay spectroscopy, and decomposition in buffers.^[149]

The development of computational approaches aiming to tackle this problem has been slow. Until recently, tools accessible to users included several rule sets, few similarity-based approaches, and a statistical approach. Among the rule sets, the best known and most applied collection is the pan-assay interference compounds (PAINS) rule set.^[149,150] Although clearly declared by its inventors, users of the PAINS rules set all too often neglect the significant limitations of its scope, applicability and reliability. Further examples of relevant rule sets include the REOS rules^[151] and a set of rules derived from an NMR-based method for identifying small molecules that cause false-positive assay outcomes due to reactivity (ALARM NMR).^[152]

A useful similarity-based approach is Aggregator Advisor, which flags compounds which are in a close structural relationship to known aggregators (a simple approach of which negative outcomes of course do not indicate the benignity of compounds).^[153] The statistical approach, called BADAPPLE,^[154] calculates a promiscuity score based on molecular scaffolds.

More recently, we introduced Hit Dexter 2.0, the second generation of a set of machine learning models that are designed to identify compounds that are likely to show frequent hitter behavior in primary screening assays and/or confirmatory dose-response assays, regardless of the underlying (interference) mechanism.^[155]

All these approaches have in common that they are derived from datasets dominated by synthetic compounds. As we point out in our work on Hit Dexter 2.0, the training set, even though consisting of about 250k compounds, covers only a small fraction (approximately 15%) of the known NPs with compounds that are structurally sufficiently similar so that reliable predictions by the model can be expected.^[155] This means, once again, that caution must be exercised when using any of these approaches in particular in the context of NPs.

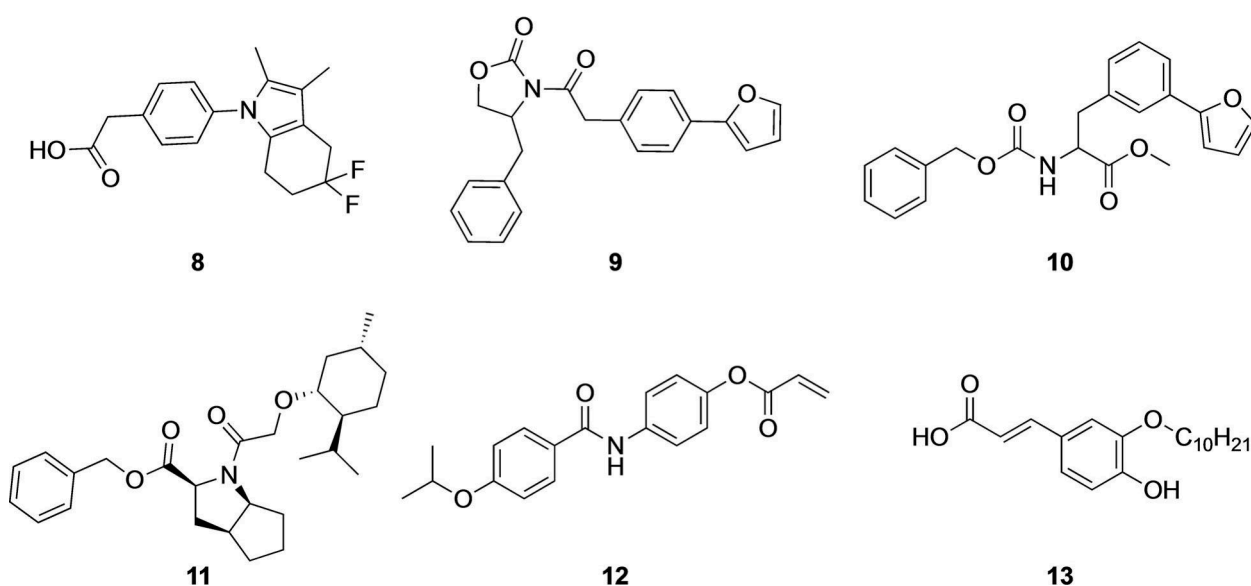


Figure 4. Examples of de novo designed molecules inspired by natural products.

10 De Novo Design of Nature-inspired Compounds and Compound Collections

Limited synthetic accessibility poses a major challenge to the exploration and use of NPs and NP-derived compounds.^[19,156] In order to overcome this hurdle, researchers have devised a number of strategies for the design of synthetically accessible compounds with NP-like properties. For example, diversity-oriented synthesis (DOS) is a concept that utilizes pairs of complexity-generating reactions to produce diverse and complex compounds with NP-like architectures (enriched with stereogenic centers and sp^3 -hybridized atoms).^[156,157] In contrast to DOS, biology-oriented synthesis (BIOS) starts from biologically active scaffolds and seeks to generate small to medium-sized collections of complexity-reduced, NP-like compounds.^[80,158] BIOS is guided by the hierarchical representation and classification of molecular scaffolds, as well as the structural similarity of the ligand-sensing cores of proteins.^[81,159]

A further strategy for the efficient synthesis of diverse, NP-like compounds utilizes chemoselective reactions for the distortion of ring systems that are part of readily available NPs.^[160,161] Common conversions in this context include ring cleavage, ring expansion, ring fusion and ring rearrangements.

Novel classes of compounds can also be derived by fragment-based compound design starting from NP-derived fragments.^[156] This NP-inspired strategy may enable the efficient exploration of the biologically relevant chemical space beyond the known NPs and NP scaffolds.

Shifting the focus to computational approaches, Hartenfeller et al.^[162] developed DOGS, a de novo design tool which utilizes information on more than 25k readily available synthetic building blocks in combination with a

large set of established reaction rules to generate compounds which are likely synthetically accessible. Importantly, DOGS utilizes structural and pharmacophoric descriptions of (bioactive) reference compounds in order to steer the compound generation process into desired directions.

Starting from NPs active on the retinoid X receptor (RXR), DOGS was employed for the design of novel, synthetically accessible, NP-inspired RXR ligands. Five out of six compounds designed by DOGS proved to be RXR agonists and to have similar nuclear receptor selectivity profiles to the respective templates (one example is **8**, shown in Figure 4).^[135] In a further study, DOGS was utilized for the design of mimics of (–)-englerin, a complex sesquiterpene with potent anti-proliferative activity.^[163] A total of 323 unique designs were generated by DOGS. After several filtering and scoring steps, two proposed molecules (**9** and **10**) were selected and synthesized (one thereof with a slight modification). Both compounds were confirmed in a functional, cell-based assay as potent inhibitors of the transient receptor potential melastatin 8 (TRPM8) ion channel.^[164]

In a follow-up study, the above-mentioned ranking approach was extended to take into account also the 3D molecular shape similarity (based on global fractal dimensionality) of the 323 designs.^[165] One of two compounds selected by this approach (**11**) was again confirmed as potent inhibitor of TRPC4 and TRPM8 channels.

Merk et al. used a deep recurrent neural network approach for the de novo design of RXR modulators.^[166] The neural network was trained on synthetic compounds with measured bioactivities on RXR. By fine-tuning the model with a small set of NPs modulators of RXR, the authors showed that their model was able to produce synthetically

accessible NP mimetics that have a high chance of being active on the intended target. Following a selection procedure that involved target prediction and the assessment of molecular similarity, three designs were selected for experimental testing of which two compounds (12 and 13) were confirmed to modulate the RXR with a potency that is comparable with that of the templates.

For additional information on de novo design in the context of NP research, the reader is referred to ref. [19].

11 Computational Prediction of ADME and Safety Profiles of Natural Products

NP-based drug discovery often faces challenges related to the ADME and safety profiles of NPs. Among the most prominent examples of anti-targets addressed by NPs is the hERG channel^[167] (its blockage is linked to potentially fatal cardiac arrhythmia), cytochrome P450 enzymes (which can cause drug-drug interactions and toxicity), and the P-glycoprotein (an efflux pump with broad substrate specificity that can effectively cause drug resistance). A plethora of computational models of different kinds (i.e. statistical models, machine learning models, pharmacophore models, docking, etc.) address these and many other anti-targets and endpoints.^[96,168–173] However, it is important to consider that, as a result of the available data, these and most other in silico models are trained and/or tested on compounds that are primarily of synthetic origin. Therefore, extra caution must be exercised in relation with NPs, and the applicability domain of the models must be closely observed.

Not all models are equally affected by the structural and physicochemical differences of NPs and synthetic compounds. For example, the applicability of Hit Dexter 2.0 to NPs is limited. The reliability of Hit Dexter's predictions has been shown to decrease substantially when moving away from the training data beyond a certain point, and the training data are primarily composed of synthetic compounds. In contrast, a conceptually related machine learning model for the prediction of the sites of metabolism of small molecules, FAME 3, was shown to perform well on NPs, even though the majority of compounds in the training set are again of synthetic origin.^[174] The reason for the high robustness of the FAME 3 models and their good performance on NPs is that the liability of atom positions in molecules is described based on their proximate atom environment, and these proximate neighborhoods are much more redundant among NPs and synthetic compounds than their global molecular similarity.

12 Summary

NPs pose some extraordinary challenges to experimentalists and theoreticians alike, but statistics on recently approved,

small-molecule medicines show that the research of NPs is worth all the effort and can yield valuable, innovative drugs. Modern in silico methods can make a substantial contribution to the acceleration and de-risking of NP-based drug discovery. However, the applicability of models must be closely observed, in particular when working with NPs as computational approaches are mostly designed for, and trained on, data for synthetic compounds. Unfortunately, even the recently developed models still often lack robust definitions of the applicability domain and do not warn users adequately about compounds for which predictions are not reliable. Researchers may in particular feel tempted to use one of the many free, user-friendly web servers to quickly predict physicochemical or biological properties of NPs. Obviously, also for these web services the principle holds true that in the absence of robust indicators of the reliability of individual predictions, these predictions are not to be trusted.

Given the reinvigorate interest in NP research, the growing amount of accessible biological, chemical and structural data, and advances in algorithms, modeling techniques and computational power, the future will see the continued integration of computational methods in NP-based drug discovery pipelines.

Conflict of interest

We declare no conflicts of interest.

Acknowledgements

Y.C. is supported by the China Scholarship Council, grant number 201606010345.

References

- [1] D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2020**, *83*, 770–803.
- [2] G. M. Cragg, D. J. Newman, *Pure Appl. Chem.* **2005**, *77*, 7–24.
- [3] T. Rodrigues, D. Reker, P. Schneider, G. Schneider, *Nat. Chem.* **2016**, *8*, 531–541.
- [4] A. G. Atanasov, B. Waltenberger, E.-M. Pferschy-Wenzig, T. Linder, C. Wawrosch, P. Uhrin, V. Temml, L. Wang, S. Schwaiger, E. H. Heiss, et al., *Biotechnol. Adv.* **2015**, *33*, 1582–1614.
- [5] J. Gu, Y. Gui, L. Chen, G. Yuan, H.-Z. Lu, X. Xu, *PLoS One* **2013**, *8*, e62839.
- [6] Y. Chen, M. G. de Lomana, N.-O. Friedrich, J. Kirchmair, *J. Chem. Inf. Model.* **2018**, *58*, 1518–1532.
- [7] P. A. Clemons, N. E. Bodycombe, H. A. Carrinski, J. A. Wilson, A. F. Shamji, B. K. Wagner, A. N. Koehler, S. L. Schreiber, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 18787–18792.
- [8] H. Chen, O. Engkvist, N. Blomberg, J. Li, *MedChemComm* **2012**, *3*, 312–321.

- [9] B. David, A. Grondin, P. Schambel, M. Vitorino, D. Zeyer, *Phytochem. Rev.* **2019**, DOI 10.1007/s11101-019-09612-4.
- [10] N.-O. Friedrich, F. Flachsenberg, A. Meyder, K. Sommer, J. Kirchmair, M. Rarey, *J. Chem. Inf. Model.* **2019**, *59*, 731–742.
- [11] N.-O. Friedrich, C. de Bruyn Kops, F. Flachsenberg, K. Sommer, M. Rarey, J. Kirchmair, *J. Chem. Inf. Model.* **2017**, *57*, 2719–2728.
- [12] N.-O. Friedrich, A. Meyder, C. de Bruyn Kops, K. Sommer, F. Flachsenberg, M. Rarey, J. Kirchmair, *J. Chem. Inf. Model.* **2017**, *57*, 529–539.
- [13] A. N. Jain, A. E. Cleves, Q. Gao, X. Wang, Y. Liu, E. C. Sherer, M. Y. Reibarkh, *J. Comput.-Aided Mol. Des.* **2019**, *33*, 531–558.
- [14] S. Wang, J. Witek, G. A. Landrum, S. Riniker, *J. Chem. Inf. Model.* **2020**, *60*, 2044–2058.
- [15] V. Poongavanam, E. Danelius, S. Peintner, L. Alcaraz, G. Caron, M. D. Cummings, S. Wlodek, M. Erdelyi, P. C. D. Hawkins, G. Ermondi, et al., *ACS Omega* **2018**, *3*, 11742–11757.
- [16] A. L. Harvey, R. Edrada-Ebel, R. J. Quinn, *Nature Rev. Drug Discov.* **2015**, *14*, 111–129.
- [17] C. J. Henrich, J. A. Beutler, *Nat. Prod. Rep.* **2013**, *30*, 1284–1298.
- [18] A. Olğaç, I. E. Orhan, E. Banoglu, *Future Med. Chem.* **2017**, *9*, 1665–1686.
- [19] T. Rodrigues, *Org. Biomol. Chem.* **2017**, *15*, 9275–9282.
- [20] N. K. K. Ikram, J. D. Durrant, M. Muchtaridi, A. S. Zalaludin, N. Purwitasari, N. Mohamed, A. S. A. Rahim, C. K. Lam, Y. M. Normi, N. A. Rahman, et al., *J. Chem. Inf. Model.* **2015**, *55*, 308–316.
- [21] U. Grienke, J. Mihály-Bison, D. Schuster, T. Afonyushkin, M. Binder, S.-H. Guan, C.-R. Cheng, G. Wolber, H. Stuppner, D.-A. Guo, et al., *Bioorg. Med. Chem.* **2011**, *19*, 6779–6791.
- [22] J. M. Rollinger, D. V. Kratschmar, D. Schuster, P. H. Pfisterer, C. Gumy, E. M. Aubry, S. Brandstötter, H. Stuppner, G. Wolber, A. Odermatt, *Bioorg. Med. Chem.* **2010**, *18*, 1507–1515.
- [23] U. Grienke, H. Braun, N. Seidel, J. Kirchmair, M. Richter, A. Krumbholz, S. von Grafenstein, K. R. Liedl, M. Schmidtke, J. M. Rollinger, *J. Nat. Prod.* **2014**, *77*, 563–570.
- [24] G. Landrum, “RDKit,” can be found under www.rdkit.org.
- [25] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
- [26] “The Chemistry Development Kit,” can be found under <https://github.com/cdk>.
- [27] “KNIME | Open for Innovation,” can be found under <https://www.knime.com/>.
- [28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- [29] “scikit-learn: machine learning in Python,” can be found under www.scikit-learn.org.
- [30] Y. Chen, C. de Bruyn Kops, J. Kirchmair, *Prog. Chem. Org. Nat. Prod.* **2019**, *110*, 37–71.
- [31] Y. Chen, C. de Bruyn Kops, J. Kirchmair, *J. Chem. Inf. Model.* **2017**, *57*, 2099–2111.
- [32] P. Banerjee, J. Erehman, B.-O. Gohlke, T. Wilhelm, R. Preissner, M. Dunkel, *Nucleic Acids Res.* **2014**, *43*, D935.
- [33] C. Y.-C. Chen, *PLoS One* **2011**, *6*, e15939.
- [34] “Natural Products Atlas (2019),” can be found under <https://www.npatlas.org>.
- [35] J. A. van Santen, G. Jacob, A. L. Singh, V. Aniebok, M. J. Balunas, D. Bunsko, F. C. Neto, L. Castaño-Espriu, C. Chang, T. N. Clark, et al., *ACS Cent. Sci.* **2019**, *5*, 1824–1833.
- [36] X. Zeng, P. Zhang, Y. Wang, C. Qin, S. Chen, W. He, L. Tao, Y. Tan, D. Gao, B. Wang, et al., *Nucleic Acids Res.* **2019**, *47*, D1118.
- [37] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, et al., *Nucleic Acids Res.* **2014**, *42*, D1083–90.
- [38] “ChEMBL Database version 23,” can be found under <https://www.ebi.ac.uk/chembl/>.
- [39] G. Wolber, T. Langer, *J. Chem. Inf. Model.* **2005**, *45*, 160–169.
- [40] “Docking files Search,” can be found under <http://docking.umh.es/chemlib/mnplib>.
- [41] M. Sorokina, C. Steinbeck, *J. Cheminf.* **2020**, *12*, 629.
- [42] T.-H. Nguyen-Vo, L. Nguyen, N. Do, T.-N. Nguyen, K. Trinh, H. Cao, L. Le, *J. Chem. Inf. Model.* **2020**, *60*, 1101–1110.
- [43] B. Yang, J. Mao, B. Gao, X. Lu, *Curr. Pharm. Biotechnol.* **2019**, *20*, 293–301.
- [44] F. Pereira, J. Aires-de-Sousa, *Mar. Drugs* **2018**, *16*, 236.
- [45] E. Koulouridi, M. Valli, F. Ntie-Kang, V. da Silva Bolzani, *Phys. Sci. Rev.* **2019**, *4*, DOI 10.1515/psr-2018-0105.
- [46] T. Sterling, J. J. Irwin, *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
- [47] “ZINC,” can be found under <http://zinc15.docking.org>.
- [48] A. Mohamed, C. H. Nguyen, H. Mamitsuka, *Briefings Bioinf.* **2016**, *17*, 309–321.
- [49] S. Chanana, C. Thomas, D. Braun, Y. Hou, T. Wyche, T. Bugni, *Metabolites* **2017**, *7*, 34.
- [50] U. Abdelmohsen, C. Cheng, C. Viegelmann, T. Zhang, T. Grkovic, S. Ahmed, R. Quinn, U. Hentschel, R. Edrada-Ebel, *Mar. Drugs* **2014**, *12*, 1220–1244.
- [51] D. C. Burns, E. P. Mazzola, W. F. Reynolds, *Nat. Prod. Rep.* **2019**, *36*, 919–933.
- [52] S. H. Martínez-Treviño, V. Uc-Cetina, M. A. Fernández-Herrera, G. Merino, *J. Chem. Inf. Model.* **2020**, *60*, 3376–3386.
- [53] R. Reher, H. W. Kim, C. Zhang, H. H. Mao, M. Wang, L.-F. Nothias, A. M. Caraballo-Rodríguez, E. Glukhov, B. Teke, T. Leao, et al., *J. Am. Chem. Soc.* **2020**, *142*, 4114–4120.
- [54] Y.-C. Harn, B.-H. Su, Y.-L. Ku, O. A. Lin, C.-F. Chou, Y. J. Tseng, *J. Chem. Inf. Model.* **2017**, *57*, 3138–3148.
- [55] I. Pérez-Victoria, J. Martín, F. Reyes, *Planta Med.* **2016**, *82*, 857–871.
- [56] S. P. Gaudêncio, F. Pereira, *Nat. Prod. Rep.* **2015**, *32*, 779–810.
- [57] P. Ertl, A. Schuffenhauer, *Prog. Drug Res.* **2008**, *66*, 217, 219–35.
- [58] S. B. Singh, J. C. Culberson, *Natural Product Chemistry for Drug Discovery* (Eds.: A. D. Buss, M. S. Butler), **2009**, pp. 28–43.
- [59] Y. Chen, C. Stork, S. Hirte, J. Kirchmair, *Biomolecules* **2019**, *9*, 43.
- [60] C. F. Stratton, D. J. Newman, D. S. Tan, *Bioorg. Med. Chem. Lett.* **2015**, *25*, 4802–4807.
- [61] D.-L. Ma, D. S.-H. Chan, C.-H. Leung, *Chem. Sci.* **2011**, *2*, 1656–1665.
- [62] M. Feher, J. M. Schmidt, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.
- [63] S. Wetzel, A. Schuffenhauer, S. Roggo, P. Ertl, H. Waldmann, *CHIMIA Int. J. Chem.* **2007**, *61*, 355–360.
- [64] F. López-Vallejo, M. A. Giulianotti, R. A. Houghten, J. L. Medina-Franco, *Drug Discovery Today* **2012**, *17*, 718–726.
- [65] K. Grabowski, G. Schneider, *Curr. Chem. Biol.* **2007**, *1*, 115–127.
- [66] K. Grabowski, K.-H. Baringhaus, G. Schneider, *Nat. Prod. Rep.* **2008**, *25*, 892–904.
- [67] T. Henkel, R. M. Brunne, H. Müller, F. Reichel, *Angew. Chem. Int. Ed.* **1999**, *38*, 643–647; *Angew. Chem.* **1999**, *111*, 688–691.
- [68] X. Lucas, B. A. Grüning, S. Bleher, S. Günther, *J. Chem. Inf. Model.* **2015**, *55*, 915–924.
- [69] J. Hert, J. J. Irwin, C. Laggner, M. J. Keiser, B. K. Shoichet, *Nat. Chem. Biol.* **2009**, *5*, 479–483.

- [70] T. El-Elmat, X. Zhang, D. Jarjoura, F. J. Moy, J. Orjala, A. D. Kinghorn, C. J. Pearce, N. H. Oberlies, *ACS Med. Chem. Lett.* **2012**, *3*, 645–649.
- [71] P. Muigg, J. Rosén, L. Bohlin, A. Backlund, *Phytochem. Rev.* **2013**, *12*, 449–457.
- [72] F. I. Saldívar-González, M. Valli, A. D. Andricopulo, V. da Silva Bolzani, J. L. Medina-Franco, *J. Chem. Inf. Model.* **2019**, *59*, 74–85.
- [73] L. I. Pilkington, *Molecules* **2019**, *24*, 3942.
- [74] J. Shang, B. Hu, J. Wang, F. Zhu, Y. Kang, D. Li, H. Sun, D.-X. Kong, T. Hou, *J. Chem. Inf. Model.* **2018**, *58*, 1182–1193.
- [75] P. Ertl, T. Schuhmann, *Mol. Inf.* **2020**, *39*, 2000017.
- [76] P. Ertl, T. Schuhmann, *J. Nat. Prod.* **2019**, *82*, 1258–1263.
- [77] A. L. Chávez-Hernández, N. Sánchez-Cruz, J. L. Medina-Franco, *Mol. Inf.* **2020**, *39*, 2000050.
- [78] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.
- [79] T. Schäfer, N. Kriege, L. Humbeck, K. Klein, O. Koch, P. Mutzel, *J. Cheminf.* **2017**, *9*, 28.
- [80] H. Lachance, S. Wetzel, K. Kumar, H. Waldmann, *J. Med. Chem.* **2012**, *55*, 5989–6001.
- [81] M. A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, H. Waldmann, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17272–17277.
- [82] M. Shen, S. Tian, Y. Li, Q. Li, X. Xu, J. Wang, T. Hou, *J. Cheminf.* **2012**, *4*, 31.
- [83] F. I. Saldívar-González, B. Angélica Pilón-Jiménez, J. L. Medina-Franco, *Phys. Sci. Rev.* **2019**, *4*, 20180103.
- [84] T. I. Oprea, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 157–166.
- [85] J. Larsson, J. Gottfries, S. Muresan, A. Backlund, *J. Nat. Prod.* **2007**, *70*, 789–794.
- [86] R. Frédéric, C. Bruyère, C. Vancaeynest, J. Reniers, C. Meinguet, L. Pochet, A. Backlund, B. Masereel, R. Kiss, J. Wouters, *J. Med. Chem.* **2012**, *55*, 6489–6501.
- [87] J. Rosén, L. Rickardson, A. Backlund, J. Gullbo, L. Bohlin, R. Larsson, J. Gottfries, *QSAR Comb. Sci.* **2009**, *28*, 436–446.
- [88] M. Korinek, Y.-H. Tsai, M. El-Shazly, K.-H. Lai, A. Backlund, S.-F. Wu, W.-C. Lai, T.-Y. Wu, S.-L. Chen, Y.-C. Wu, et al., *Front. Pharmacol.* **2017**, *8*, 356.
- [89] M. Pascolutti, M. Campitelli, B. Nguyen, N. Pham, A.-D. Gorse, R. J. Quinn, *PLoS One* **2015**, *10*, e0120942.
- [90] T. Miyao, D. Reker, P. Schneider, K. Funatsu, G. Schneider, *Planta Med.* **2015**, *81*, 429–435.
- [91] L. van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
- [92] L. McInnes, J. Healy, J. Melville, *arXiv e-prints* **2018**, 1802.03426v2.
- [93] M. González-Medina, F. D. Prieto-Martínez, J. R. Owen, J. L. Medina-Franco, *J. Cheminf.* **2016**, *8*, 63.
- [94] M. González-Medina, J. R. Owen, T. El-Elmat, C. J. Pearce, N. H. Oberlies, M. Figueroa, J. L. Medina-Franco, *Front. Pharmacol.* **2017**, *8*, 180.
- [95] D. A. Olmedo, M. González-Medina, M. P. Gupta, J. L. Medina-Franco, *Mol. Diversity* **2017**, *21*, 779–789.
- [96] F. D. Prieto-Martínez, U. Norinder, J. L. Medina-Franco, *Prog. Chem. Org. Nat. Prod.* **2019**, *110*, 1–35.
- [97] N. Sánchez-Cruz, J. L. Medina-Franco, *J. Cheminf.* **2018**, *10*, 55.
- [98] M. J. Yu, *J. Chem. Inf. Model.* **2011**, *51*, 541–557.
- [99] P. Ertl, S. Roggo, A. Schuffenhauer, *J. Chem. Inf. Model.* **2008**, *48*, 68–74.
- [100] K. V. Jayaseelan, P. Moreno, A. Truszkowski, P. Ertl, C. Steinbeck, *BMC Bioinf.* **2012**, *13*, 106.
- [101] K. V. Jayaseelan, C. Steinbeck, *BMC Bioinf.* **2014**, *15*, 234.
- [102] “RDKit NP_Score,” can be found under https://github.com/rdkit/rdkit/tree/master/Contrib/NP_Score.
- [103] M. Sorokina, C. Steinbeck, *J. Cheminf.* **2019**, *11*, 55.
- [104] H. Zaid, J. Raiyn, A. Nasser, B. Saad, A. Rayan, *Open Nutraceuticals J.* **2010**, *3*, 194–202.
- [105] “NP-Scout,” can be found under <https://nerdd.zbh.uni-hamburg.de/npscout/>.
- [106] M. Seo, H. K. Shin, Y. Myung, S. Hwang, K. T. No, *J. Cheminf.* **2020**, *12*, 6.
- [107] B. Kirchweger, J. M. Rollinger, *Natural Products as Source of Molecules with Therapeutic Potential* (Ed.: V. C. Filho), **2018**, pp. 333–364.
- [108] B. Kirchweger, J. M. Rollinger, *Prog. Chem. Org. Nat. Prod.* **2019**, *110*, 239–271.
- [109] U. Grienke, M. Schmidtke, J. Kirchmair, K. Pfarr, P. Wutzler, R. Dürrwald, G. Wolber, K. R. Liedl, H. Stuppner, J. M. Rollinger, *J. Med. Chem.* **2010**, *53*, 778–786.
- [110] R. E. Amaro, J. Baudry, J. Chodera, Ö. Demir, J. A. McCammon, Y. Miao, J. C. Smith, *Biophys. J.* **2018**, *114*, 2271–2278.
- [111] G. L. Warren, C. W. Andrews, A.-M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, et al., *J. Med. Chem.* **2006**, *49*, 5912–5931.
- [112] T. Seidel, O. Wieder, A. Garon, T. Langer, *Mol. Inf.* **2020**, DOI 10.1002/minf.202000059.
- [113] J. Kirchmair, J. M. Rollinger, K. R. Liedl, N. Seidel, A. Krumbholz, M. Schmidtke, *Future Med. Chem.* **2011**, *3*, 437–450.
- [114] B. Kirchweger, J. M. Kratz, A. Ladurner, U. Grienke, T. Langer, V. M. Dirsch, J. M. Rollinger, *Front. Chem.* **2018**, *6*, 242.
- [115] D. Schuster, P. Markt, U. Grienke, J. Mihaly-Bison, M. Binder, S. M. Noha, J. M. Rollinger, H. Stuppner, V. N. Bochkov, G. Wolber, *Bioorg. Med. Chem.* **2011**, *19*, 7168–7180.
- [116] M. Rupp, T. Schroeter, R. Steri, H. Zettl, E. Proschak, K. Hansen, O. Rau, O. Schwarz, L. Müller-Kuhr, M. Schubert-Zsilavec, et al., *ChemMedChem* **2010**, *5*, 191–194.
- [117] F. Grisoni, D. Merk, L. Friedrich, G. Schneider, *ChemMedChem* **2019**, *14*, 1129–1134.
- [118] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, G. Pujadas, S. Garcia-Vallve, *Methods* **2015**, *71*, 98–103.
- [119] A. Ezzat, M. Wu, X.-L. Li, C.-K. Kwok, *Briefings Bioinf.* **2019**, *20*, 1337–1357.
- [120] E. Sam, P. Athri, *Briefings Bioinf.* **2019**, *20*, 299–316.
- [121] R. Chaudhari, Z. Tan, B. Huang, S. Zhang, *Expert Opin. Drug Discovery* **2017**, *12*, 279–291.
- [122] N. Mathai, Y. Chen, J. Kirchmair, *Briefings Bioinf.* **2019**, *21*, 791–802.
- [123] N. Mathai, J. Kirchmair, *Int. J. Mol. Sci.* **2020**, *21*, 3585.
- [124] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijter, R. C. Matos, T. B. Tran, et al., *Nature* **2009**, *462*, 175–181.
- [125] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, et al., *Nature* **2012**, *486*, 361–367.
- [126] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, B. K. Shoichet, *Nat. Biotechnol.* **2007**, *25*, 197–206.
- [127] D. Gfeller, A. Grosdidier, M. Wirth, A. Daina, O. Michielin, V. Zoete, *Nucleic Acids Res.* **2014**, *42*, W32–8.
- [128] “ROCS. OpenEye Scientific Software,” can be found under <https://www.eyesopen.com>.
- [129] P. C. D. Hawkins, A. G. Skillman, A. Nicholls, *J. Med. Chem.* **2007**, *50*, 74–82.
- [130] Y. Chen, N. Mathai, J. Kirchmair, *J. Chem. Inf. Model.* **2020**, *60*, 2858–2875.

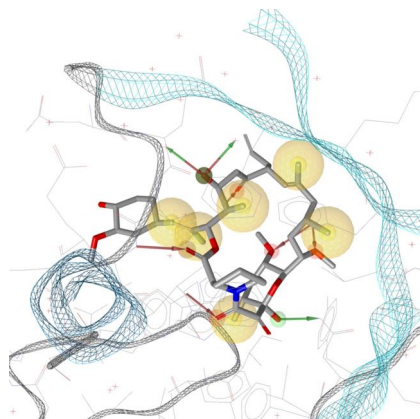
- [131] J. M. Rollinger, D. Schuster, B. Danzl, S. Schwaiger, P. Markt, M. Schmidtke, J. Gertsch, S. Raduner, G. Wolber, T. Langer, et al., *Planta Med.* **2009**, *75*, 195–204.
- [132] D. Reker, T. Rodrigues, P. Schneider, G. Schneider, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 4067–4072.
- [133] P. Schneider, G. Schneider, *Chem. Commun.* **2017**, *53*, 2272–2274.
- [134] N. T. Cockroft, X. Cheng, J. R. Fuchs, *J. Chem. Inf. Model.* **2019**, *59*, 4906–4920.
- [135] D. Merk, F. Grisoni, L. Friedrich, E. Gelzinyte, G. Schneider, *J. Med. Chem.* **2018**, *61*, 5442–5447.
- [136] T. Rodrigues, F. Sieglitz, V. J. Somovilla, P. M. S. D. Cal, A. Galione, F. Corzana, G. J. L. Bernardes, *Angew. Chem. Int. Ed. Engl.* **2016**, *55*, 11077–11081.
- [137] D. Reker, A. M. Perna, T. Rodrigues, P. Schneider, M. Reutlinger, B. Mönch, A. Koeberle, C. Lamers, M. Gabler, H. Steinmetz, et al., *Nat. Chem.* **2014**, *6*, 1072–1078.
- [138] G. Schneider, D. Reker, T. Chen, K. Hauenstein, P. Schneider, K.-H. Altmann, *Angew. Chem. Int. Ed. Engl.* **2016**, *55*, 12408–12411.
- [139] T. Rodrigues, D. Reker, J. Kunze, P. Schneider, G. Schneider, *Angew. Chem. Int. Ed. Engl.* **2015**, *54*, 10516–10520.
- [140] T. Rodrigues, M. Werner, J. Roth, E. H. G. da Cruz, M. C. Marques, P. Akkapeddi, S. A. Lobo, A. Koeberle, F. Corzana, E. N. da Silva Júnior, et al., *Chem. Sci.* **2018**, *9*, 6899–6903.
- [141] P. Schneider, G. Schneider, *Angew. Chem. Int. Ed. Engl.* **2017**, *56*, 11520–11524.
- [142] J. Fang, Z. Wu, C. Cai, Q. Wang, Y. Tang, F. Cheng, *J. Chem. Inf. Model.* **2017**, *57*, 2657–2671.
- [143] J. Fang, L. Gao, H. Ma, Q. Wu, T. Wu, J. Wu, Q. Wang, F. Cheng, *Front. Pharmacol.* **2017**, *8*, 747.
- [144] J. Lai, J. Hu, Y. Wang, X. Zhou, Y. Li, L. Zhang, Z. Liu, *Mol. Inf.* **2020**, *39*, 2000057.
- [145] D. Sydow, L. Burggraaff, A. Szengel, H. W. T. van Vlijmen, A. P. IJzerman, G. J. P. van Westen, A. Volkamer, *J. Chem. Inf. Model.* **2019**, *59*, 1728–1742.
- [146] J. Bisson, J. B. McAlpine, J. B. Friesen, S.-N. Chen, J. Graham, G. F. Pauli, *J. Med. Chem.* **2016**, *59*, 1671–1690.
- [147] F. E. Koehn, G. T. Carter, *Nat. Rev. Drug Discovery* **2005**, *4*, 206–220.
- [148] S. L. McGovern, E. Caselli, N. Grigorieff, B. K. Shoichet, *J. Med. Chem.* **2002**, *45*, 1712–1722.
- [149] J. B. Baell, G. A. Holloway, *J. Med. Chem.* **2010**, *53*, 2719–2740.
- [150] J. B. Baell, J. W. M. Nissink, *ACS Chem. Biol.* **2018**, *13*, 36–44.
- [151] W. P. Walters, M. T. Stahl, M. A. Murcko, *Drug Discovery Today* **1998**, *3*, 160–178.
- [152] J. R. Huth, R. Mendoza, E. T. Olejniczak, R. W. Johnson, D. A. Cothron, Y. Liu, C. G. Lerner, J. Chen, P. J. Hajduk, *J. Am. Chem. Soc.* **2005**, *127*, 217–224.
- [153] J. J. Irwin, D. Duan, H. Torosyan, A. K. Doak, K. T. Ziebart, T. Sterling, G. Tumanian, B. K. Shoichet, *J. Med. Chem.* **2015**, *58*, 7076–7087.
- [154] J. J. Yang, O. Ursu, C. A. Lipinski, L. A. Sklar, T. I. Oprea, C. G. Bologa, *J. Cheminf.* **2016**, *8*, 29.
- [155] C. Stork, Y. Chen, M. Šicho, J. Kirchmair, *J. Chem. Inf. Model.* **2019**, *59*, 1030–1043.
- [156] G. Karageorgis, D. J. Foley, L. Laraia, H. Waldmann, *Nat. Chem.* **2020**, *12*, 227–235.
- [157] T. E. Nielsen, S. L. Schreiber, *Angew. Chem. Int. Ed. Engl.* **2008**, *47*, 48–56.
- [158] S. Wetzel, R. S. Bon, K. Kumar, H. Waldmann, *Angew. Chem. Int. Ed. Engl.* **2011**, *50*, 10800–10826.
- [159] S. Renner, W. A. L. van Otterlo, M. Dominguez Seoane, S. Möcklinghoff, B. Hofmann, S. Wetzel, A. Schuffenhauer, P. Ertl, T. I. Oprea, D. Steinhilber, et al., *Nat. Chem. Biol.* **2009**, *5*, 585–592.
- [160] R. W. Huigens 3rd, K. C. Morrison, R. W. Hicklin, T. A. Flood Jr, M. F. Richter, P. J. Hergenrother, *Nat. Chem.* **2013**, *5*, 195–202.
- [161] R. J. Rafferty, R. W. Hicklin, K. A. Maloof, P. J. Hergenrother, *Angew. Chem. Int. Ed. Engl.* **2014**, *53*, 220–224.
- [162] M. Hartenfeller, H. Zettl, M. Walter, M. Rupp, F. Reisen, E. Proschak, S. Weggen, H. Stark, G. Schneider, *PLoS Comput. Biol.* **2012**, *8*, e1002380.
- [163] Y. Akbulut, H. J. Gaunt, K. Muraki, M. J. Ludlow, M. S. Amer, A. Bruns, N. S. Vasudev, L. Radtke, M. Willot, S. Hahn, et al., *Angew. Chem. Int. Ed. Engl.* **2015**, *54*, 3787–3791.
- [164] L. Friedrich, T. Rodrigues, C. S. Neuhaus, P. Schneider, G. Schneider, *Angew. Chem. Int. Ed. Engl.* **2016**, *55*, 6789–6792.
- [165] L. Friedrich, R. Byrne, A. Treder, I. Singh, C. Bauer, T. Gudermann, M. M. y. Schnitzler, U. Storch, G. Schneider, *ChemMedChem* **2020**, *15*, 566–570.
- [166] D. Merk, F. Grisoni, L. Friedrich, G. Schneider, *Commun. Chem.* **2018**, *1*, 68.
- [167] J. M. Kratz, U. Grienke, O. Scheel, S. A. Mann, J. M. Rollinger, *Nat. Prod. Rep.* **2017**, *34*, 957–980.
- [168] M. P. Gleeson, S. Modi, A. Bender, R. L. M. Robinson, J. Kirchmair, M. Promkatkaew, S. Hannongbua, R. C. Glen, *Curr. Pharm. Des.* **2012**, *18*, 1266–1291.
- [169] J. Kirchmair, A. H. Göller, D. Lang, J. Kunze, B. Testa, I. D. Wilson, R. C. Glen, G. Schneider, *Nat. Rev. Drug Discovery* **2015**, *14*, 387–404.
- [170] H. Yang, L. Sun, W. Li, G. Liu, Y. Tang, *Front. Chem.* **2018**, *6*, 30.
- [171] Y. Wang, J. Xing, Y. Xu, N. Zhou, J. Peng, Z. Xiong, X. Liu, X. Luo, C. Luo, K. Chen, et al., *Q. Rev. Biophys.* **2015**, *48*, 488–515.
- [172] H. K. Shin, Y.-M. Kang, K. T. No, *Handbook of Computational Chemistry* **2016**, 1–37.
- [173] C.-Y. Jia, J.-Y. Li, G.-F. Hao, G.-F. Yang, *Drug Discovery Today* **2020**, *25*, 248–258.
- [174] M. Šicho, C. Stork, A. Mazzolari, C. de Bruyn Kops, A. Pedretti, B. Testa, G. Vistoli, D. Svozil, J. Kirchmair, *J. Chem. Inf. Model.* **2019**, *59*, 3400–3412.

Received: May 12, 2020

Accepted: July 28, 2020

Published online on September 6, 2020

REVIEW



*Y. Chen, J. Kirchmair**

1 – 17

**Cheminformatics in Natural
Product-Based Drug Discovery**

1.2. Aims

A wide range of cheminformatics methods are applied in natural products-based drug discovery, but most of them are designed for, and trained on data from, synthetic compounds. The goal of this doctoral research project was to obtain an in-depth understanding of the quality, quantity and reach of the data on NPs available in the public domain, and to advance computational methods that allow the exploitation of these data, for example for virtual screening and target prediction.

The aim of the first phase of this research project is to obtain a comprehensive overview of data resources for the computer-guided discovery of bioactive NPs, including resources focused on the physicochemical properties of NPs or their interaction with macromolecules. The study should render, for the first time, a clean and comprehensive picture of the number of known NPs and readily obtainable NPs ([Chapter 2.1](#)).

With the new insights gained during this study and the cheminformatics infrastructure in place, the aim of the second phase of this research project is to obtain an in-depth understanding of the physicochemical and structural properties of NPs and the subset of readily obtainable NPs, and how these properties compare to those of approved drugs ([Chapter 3.1](#)).

Many NP data sets have quality issues. For example, while claiming to consist of only genuine NPs, many databases also contain NP-derivatives and NP-analogs. On the other hand, many compound libraries presented as synthetic compound collections are found to contain surprisingly high numbers of NPs. In order to maximize the accessibility of chemical data on NPs, the aim of the third part of this study is to devise a novel machine learning approach for the identification of NPs and NP-like compounds in large molecular libraries ([Chapter 3.2](#)).

The aim of the final part of this study is to adopt a computational method for one of the most pressing questions in cheminformatics, and in the context of NP research in particular: What are the macromolecular targets of small molecules? A few studies suggest that computational approaches can make a significant contribution to answering this question but no systematic study on this topic in the context of natural products research has been published as of yet. As we show in [Chapter 3.3](#), 3D approaches based on molecular shape similarity may be able to recognize even distant molecular similarity and to provide valuable indications of the likely targets of NPs.

2. Data Resources and Methods

In this section, the data resources and methods of direct relevance to this work are described. Additional information on cheminformatics methods relevant to natural products-based drug discovery is provided in [D1](#).

2.1. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products

Due to the importance of NPs in drug discovery an increasing number of NP databases have become available in recent years. These databases enable large-scale virtual screening for the discovery of bioactive NPs, and the study of the physicochemical and biological properties characteristic to NPs. They can provide inspiration for NP-based drug discovery and be of value to many additional applications.

In 2017, we conducted a comprehensive and detailed survey of data resources on NPs focused on resources that are of relevance to the computer-assisted discovery of bioactive NPs. In total, we reviewed 25 virtual and 31 physical NP libraries and determined the types, quantity and quality of data provided by these resources. ([D2](#))

One of the main conclusions of this work is that the number of known NPs is approximately 250k, and that roughly 10% (25k) of these NPs are readily obtainable for testing. By allowing minor structural deviations to include mainly NP derivatives and analogs, the number of readily obtainable compounds increases by roughly 10k to 30k. Furthermore, a large number of fragment-sized NPs especially for the readily purchasable NPs were identified.

[D2] Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* **2017**, *57* (9), 2099–2111.

Available at <https://doi.org/10.1021/acs.jcim.7b00341>.

Y. Chen, C. de Bruyn Kops and J. Kirchmair conceptualized the work. Y. Chen and C. de Bruyn Kops analyzed the literature and wrote the largest part of the manuscript. Y. Chen collected, curated and analyzed all the data. J. Kirchmair supervised this work.

Reprinted with permission from

Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* **2017**, *57* (9), 2099–2111.

Copyright 2017 American Chemical Society

Data Resources for the Computer-Guided Discovery of Bioactive Natural Products

Ya Chen,^{†,‡, #} Christina de Bruyn Kops,^{†,‡, #} and Johannes Kirchmair^{*, †, ‡}

[†]Center for Bioinformatics, Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, Hamburg 20146, Germany



ABSTRACT: Natural products from plants, animals, marine life, fungi, bacteria, and other organisms are an important resource for modern drug discovery. Their biological relevance and structural diversity make natural products good starting points for drug design. Natural product-based drug discovery can benefit greatly from computational approaches, which are a valuable precursor or supplementary method to *in vitro* testing. We present an overview of 25 virtual and 31 physical natural product libraries that are useful for applications in cheminformatics, in particular virtual screening. The overview includes detailed information about each library, the extent of its structural information, and the overlap between different sources of natural products. In terms of chemical structures, there is a large overlap between freely available and commercial virtual natural product libraries. Of particular interest for drug discovery is that at least ten percent of known natural products are readily purchasable and many more natural products and derivatives are available through on-demand sourcing, extraction and synthesis services. Many of the readily purchasable natural products are of small size and hence of relevance to fragment-based drug discovery. There are also an increasing number of macrocyclic natural products and derivatives becoming available for screening.

KEYWORDS: *Natural product databases, Natural products, Chemical space, Traditional Chinese medicine, Drug discovery, Virtual screening, Plants, Maritime species, Vendors, Purchasable compounds*

INTRODUCTION

Natural products (NPs) are historically and currently relevant as components of traditional medicines and herbal remedies.^{1,2} Botanicals in particular have been used worldwide throughout history to treat various afflictions, and some of the traditional healing practices involving NPs, including traditional Chinese medicine (TCM) and traditional Indian healing systems such as Ayurveda, remain the primary treatment option for many people. Other ancient civilizations, among them Mesopotamia, ancient Egypt, and ancient Greece, also documented their use of medicinal plants, animal products, and minerals. Herbal remedies were relied on during the dark and middle ages in Europe as well. The transition from natural remedies of unknown molecular content to modern western medicine began in the early 19th century with the isolation of morphine from opium, followed quickly by the isolation of further alkaloids from plants.³ More recently, the transition from traditional remedies to single-compound drugs has resulted in the development of drugs such as artemisinin, a notable success story of a drug discovered from traditional Chinese medicine

that earned its discoverer the 2015 Nobel Prize in Physiology or Medicine.^{4,5}

Natural products already play an important role in drug development. Unlike traditional medicines, which rely primarily on herbal remedies, minerals, and animal products due to their greater accessibility, modern drugs that are NPs or NP derivatives also come from marine life, fungi, bacteria, and other organisms. Today most major classes of antibiotics, from penicillins to macrolides, are based on NPs isolated from microbes,² as are two of the three main currently used classes of antifungals: polyenes and echinocandins.⁶ In addition, many other compounds used to treat various diseases are NPs or NP derivatives. Structurally, over half of all small-molecule drugs approved between 1981 and 2014 resemble NPs.¹ Of those, 6% are unaltered NPs, 26% are NP derivatives, and 32% mimic a NP and/or contain a NP pharmacophore.¹ Examples of drugs from the latter two categories, along with the corresponding

Received: June 7, 2017

Published: August 30, 2017



NPs, are shown in Figure 1. Furthermore, certain categories of drugs are based to an even higher extent on NPs. For example,

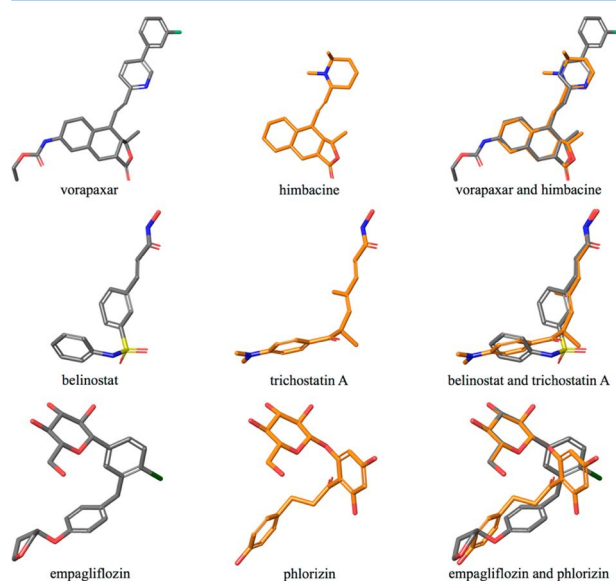


Figure 1. Three examples of a drug (left column) compared to the NP (middle column) it is derived from or mimics. A flexible alignment with ROCS⁷ is shown in the third column for each drug–NP pair (drug in gray, natural product in orange). Vorapaxar is a derivative of the NP himbacine,⁸ belinostat mimics the NP trichostatin A,^{1,9} and empagliflozin mimics and contains the pharmacophore of the NP phlorizin.^{1,10} Atom colors: oxygen in red, nitrogen in blue, sulfur in yellow, fluorine in green, and carbon in gray and orange for the drug and natural product, respectively.

73% of small-molecule antibacterials approved between 1981 and 2014 are NPs or NP derivatives.¹

The benefits of NPs as starting points for drug discovery can be explained by the biological relevance of their structures. The success of screening depends to a large extent on the structural diversity of the compound library and the pharmacological relevance of the scaffolds it contains. Meanwhile, most combinatorial screening libraries are developed around synthetic compounds and designed to be drug-like or lead-like in order to preemptively take synthesizability and oral bioavailability into account. Despite the intention of designing combinatorial libraries that cover as large a portion of the chemical space as possible, the imposed constraints of drug-likeness and lead-likeness constrict these libraries to some portion of the chemical space representing oral bioavailability rather than covering the entirety of the biologically relevant chemical space.² Even large, diverse combinatorial libraries can only encompass a small portion of all possible structural and chemical diversity.¹¹

Natural products and NP-derived drugs have more physicochemical and structural diversity than synthetic drugs and also represent a larger area of the chemical space.¹² Many NPs, however, would not be considered drug-like or lead-like based on typical physicochemical properties.¹³ Yet NPs stem from an organism's adaptation to its environment and thus have an evolution-based, specific biological purpose. It therefore makes sense that NPs generally contain biologically relevant scaffolds.¹⁴ Substructures or scaffolds of NPs are often considered privileged structures,^{15–17} meaning they may offer

improved bioactivity against diverse targets.¹⁸ One study has found that nearly 1300 ring scaffolds (with up to 11 heavy atoms in the molecule) found in NPs, or 83% of all such ring scaffolds, are not found in commercial compound libraries.¹⁹ This finding indicates the potential of NP structural features that is untapped in terms of combinatorial libraries. Hence it is clear that the benefits to using NPs in drug design are easily overlooked by typical combinatorial screening libraries.^{20,21} Compound libraries can, however, be created in order to take advantage of the opportunities presented by NPs and their scaffolds. Such a library could be composed of NPs that have been filtered for drug- or lead-likeness or based on NP scaffolds within small molecules that are easier to synthesize than larger, more structurally complex NPs.^{2,17} NPs can also be used as starting points for modifications, such as site- and stereo-selective transformations as well as the addition of reactive groups, to further increase the diversity of NP-based libraries.²²

Natural products and their derivatives have been shown to have higher hit rates in high-throughput screening (HTS) than traditional synthetic combinatorial libraries;^{20,21} however, NPs present difficulties for HTS that synthetic compound libraries do not. Isolated NPs are rarely available in sufficient quantities for HTS use,¹¹ and running HTS on isolated NPs results in an artificial bias toward those available in ample abundance. Therefore, HTS is typically carried out using crude extracts from organisms, and the active component(s) must be subsequently isolated and characterized. Despite recent advances in HTS specifically in the context of crude NP extracts, such as prefractionation and choice of a favorable assay type, several complications remain.¹¹ These obstacles to successful HTS stem mostly from the multicomponent nature of the extract, as some compounds may interfere with the assay as well as decompose, aggregate or precipitate.³ In addition, the assay result for a crude extract may be confounded by antagonistic or synergistic interactions among its components.²³ In this case, different experimental results would be achieved with isolated constituents. A similar problem arises from widely varying concentrations of the compounds contained in a crude extract, varying from too low for detection to so high that nonspecific inhibition confounds the results.²³ Added to these difficulties with HTS are obstacles such as the extraction and characterization of a sample, followed by the extensive time and effort necessary to isolate a NP of interest as well as to synthesize, partially synthesize, or modify it.

Computational approaches enable saving time and resources by identifying more promising compounds and focusing the effort of extraction, purification, or synthesis on the NPs with the most encouraging results from computational methods such as docking, quantitative structure–activity relationship (QSAR) modeling, and pharmacophore-based methods. Computational approaches can be used, for example, to predict binding affinities to a particular target, to predict ADME (absorption, distribution, metabolism, excretion) and toxicological properties, and to elucidate the biological significance of an observed effect, such as that of an herbal remedy. To make useful NP-related predictions with computational approaches, however, it is necessary to have access to accurate structures with defined stereochemistry.

When using virtual screening for NPs, the higher structural complexity and larger portion of potentially undesirable or reactive substructures compared to synthetic, drug-like molecules should be taken into account. NPs tend to have more chiral centers than drug-like, synthetic compounds.^{12,23,24}

Table 1. Virtual Natural Product Libraries

Library	No. of molecules ^a	Bioactivities ^b	Free use ^c	Molecular libraries provided free of charge ^d	Chemistry-aware Web interface ^e	Included in the analysis ^f	Scientific literature	Online presence
Dictionary of Natural Products (DNP)	>230k (>153k)	yes	no	no	yes	yes		38
Reaxys	>220k	yes	no	no	yes	no		39
Super Natural II	>325k	yes	yes	no	yes	no	40	41
UNPD	>229k (>167k)	no	yes	yes	no	yes	42	43
TCM database@ Taiwan	>60k (~50k)	yes (traditional Chinese medicines)	yes	yes	yes	yes	44	45
TCMID	>13k (>11k)	yes (traditional Chinese medicines)	yes	yes	no	yes	46	47
Chem-TCM	>12k	yes	no	no	yes	no	48	49
HIT	>700 (>400)	yes	yes	via ZINC	yes	yes	50	51
HIM	~1300 (~700)	yes (focus on ADME and toxicity data)	yes	via ZINC	yes	yes	52	53
AfroDb	~1000 (~900)	yes	yes	yes	no	yes	54	55
AfroCancer	~400 (>350)	yes (focus on anticancer activity)	yes	yes	no	yes	56	57
AfroMalariaDB	>250 (~250)	yes (focus on antimalaria activity)	yes	yes	no	yes	58	59
SANCDDB	>600 (~600)	no	yes	yes	yes	yes	60	61
NANPDB	>4400 (~3900)	yes	yes	yes	yes	yes	62	63
NPACT	~1500 (~1400)	yes (focus on anticancer activity)	yes	via ZINC	yes	yes	64	65
NPCARE	>6500 from online search; >1500 in bulk download (>1500)	yes (focus on anticancer activity)	yes	yes	no	yes	66	67
TIPdb	~9000 (~8000)	yes (focus on anticancer, antiplatelet and antituberculosis activity)	yes	yes	no	yes	68,69	70
Natural Products in PubChem Substance Database	~3000 (~2800)	yes	yes	yes	yes	yes	71	72
StreptomeDB	~4000 (~3600)	yes	yes	yes	yes	yes	73	74
UEFS Natural Products	~500 (~500)	no	via ZINC	via ZINC	no	yes		
NuBBE database	>1800, including >1700 plant NPs and >100 microorganism NPs (~1700)	yes (focus on antimicrobial activity)	yes	yes	yes	yes	75	76
Carotenoids Database	>1100	yes	yes	no	yes	no	77	78
AntiBase	>40k	yes	no	no	yes	no	79	80
DMNP	>55k (including NP derivatives)	yes	no	no	yes	no		81
MarinLit	>29k	yes	no	no	yes	no		82

^aNumber of molecules reported in the primary literature, on the Web site of the database provider, or supplied in the original files. The number in brackets reports the number of unique molecules as defined by unique InChIs^{83,84} (without the stereochemistry and fixed hydrogen layers) among the sets of standardized molecules (counterions of salts removed and compounds neutralized with the Wash function in MOE⁸⁵) for any data sets accessible to the authors. ^bIndicates whether a database includes bioactivity data that can be downloaded or accessed via a Web interface. ^cIndicates whether a database can be used free of charge, either via download or a Web interface. ^dIndicates whether the molecular structures of a database are downloadable in bulk or available upon request from the authors free of charge. ^eIndicates whether a chemistry-aware Web interface and search functionality (such as exact structure, substructure and similarity search) is provided. ^fIndicates whether a database has been included in the analysis presented in this review.

Unknown chirality may make it necessary to enumerate and screen all possible stereochemical configurations of a particular molecule in order to predict whether such a compound could bind to the target at all. If the chirality is unknown or undefined for many atoms, such a process is costly and, in any case, the actual chirality of the NP may remain undetermined. The stereochemistry can only be definitively determined experimentally, after extraction and purification of the NP.

In addition to higher stereochemical complexity, NPs as a whole have higher shape complexity, i.e., more complex scaffolds.²⁵ These scaffolds are often made up of larger, more complex ring systems, including fused ring systems that contribute to more rigid scaffolds. Further, NPs contain a larger portion of sp³-hybridized bridgehead atoms than

synthetic small molecules^{26,27} and generally tend to be less aromatic, with arene systems contained in only 38% of known NPs.¹⁷ Interestingly, this higher 3D shape complexity exists concurrently with the trend of privileged substructures in NPs. Of greater concern for virtual screening purposes is that undesirable functional groups and substructures are more prevalent in NPs, as well as that NPs tend to have more reactive elements.^{17,23} Nearly half of all known NPs contain a reactive or undesired substructure, in contrast to less than 10% of approved drugs.¹⁷ Multitarget interactions and target class promiscuity, however, are more likely to be found in synthetic compounds than in NPs.^{17,25} In terms of undesired substructures, glycosides and molecules with other metabolically unstable functional groups such as esters are likely quickly

cleaved in vivo. The core structure can be better taken into account by using SMARTS patterns or other algorithms to remove such moieties prior to virtual screening (see, e.g., ref 28). A more subtle difficulty regarding virtual screening with NPs is that any force fields and force field-based approaches, as well as computational models for predicting physicochemical and biological properties, are in general biased toward synthetic molecules. This effect has arisen from the larger abundance of synthetic molecules combined with their lower complexity compared to NPs. Despite these difficulties, virtual screening of NPs can lead to the discovery of compounds with promising in vitro or in vivo activity (see, e.g., refs 29–31).

Virtual NP libraries provide a bridge between virtual screening and the benefits of NPs for drug discovery. Screening libraries of known NPs can circumvent problems with extracting and purifying samples, or at least postpone these difficulties until later, when they can be directed toward only the most promising NPs. Previous reviews of NP databases (e.g., refs 32–37) have concentrated on a general comparison of several databases. For example, Füllbeck et al.³³ describe five publically available virtual databases containing NPs, six commercially available virtual NP databases, and companies and suppliers of NPs broken down by part of the world. On the other hand, Lagunin et al.³⁵ and Tung³⁴ focus on databases of plant natural products. In addition, other NP reviews have provided a short list of NP databases that can be used for virtual screening (e.g., refs 2,23).

In contrast to these previous publications, the focus of this review is on NP libraries that are useful for applications in cheminformatics, specifically virtual screening. We provide a detailed discussion of available virtual and physical NP libraries, along with a comparison of their contents. We additionally consider the portion of virtually available NPs that are readily purchasable.

■ VIRTUAL NATURAL PRODUCT LIBRARIES

There are many virtual NP databases in existence, both commercial and freely available. These databases vary in size, focus, and the types of information they contain for each compound. Here we focus on the databases that can be of use for virtual screening and further applications in cheminformatics. In particular, we prioritize downloadability of chemical structures with annotated stereochemistry. If the complete database is not available for download either online or upon request, then we require, at a minimum, chemistry-aware search functionality on the Web site with access to the chemical structures of the search results. The information provided here is not intended as an exhaustive list of all NP databases but rather as an overview of those that are of particular interest for virtual screening (Table 1).

Comprehensive Databases. *Dictionary of Natural Products (DNP)*. The Dictionary of Natural Products (DNP)^{38,86} is one of the most comprehensive collections of NPs available to date. This commercial database includes information on names and synonyms, physicochemical properties (e.g., molecular weight, pK_a , solubilities and spectroscopic data) and molecular structure of NPs, in addition to biological source and use. Stereochemistry for the structures is only included as a property and indicated in Fisher-type diagrams, separate from the 2D connection tables and InChIs. The compounds are classified into structural type, with a total of over 1050 classes. An abundance of physicochemical and structural information is provided, including UV spectra,

biological sources, hazards, toxicity data, and dissociation constants. The database is available both online and as CD-ROM.

Reaxys. A commercial chemical database with a focus on providing detailed information for synthetic chemists, Reaxys³⁹ contains extensive information on over 220k NPs collected from a large selection of periodicals.⁸⁷ The information on the NPs contained in Reaxys includes structures, reactions, physical properties, biological sources, and bioactivity data. The NPs can be accessed via the Web interface, which provides detailed search functionality. It is possible to search for all NPs and then download the search results in SD or SMILES file format.

Super Natural II. Super Natural II, with over 325k natural compounds, is currently the largest freely available database of NPs. This database provides chemical structures, physicochemical properties and predicted toxicity classes.⁴⁰ The compounds were collected from 16 vendor databases and five freely available databases (KEGG,⁸⁸ MetaCyc,⁸⁹ UNPD,⁴² HMDB,⁹⁰ and ZINC⁹¹). The stereochemistry is defined via chirality flags in the MOL files and isomeric SMILES. All structures and corresponding information are available for download in the form of one individual MOL file per compound from the Web site; however, it is not possible to download the entire database at once. The Web interface allows structure search (substructure and similarity) as well as searching by compound classification, several physicochemical properties, and supplier. The Web site additionally provides a search functionality for the mechanism of action starting with either a molecular (sub-) structure or a target.

UNPD. The Universal Natural Products Database (UNPD) is the largest freely available NP database that can be downloaded in full.⁴² Currently containing over 229k compounds, this database is a consolidation of NPs from several preexisting databases: Reaxys,³⁹ Chinese Natural Product Database (CNPD),⁹² CHDD⁹³ (database containing components of Chinese traditional medicinal herbs, previously developed by the authors of the UNPD), and Traditional Chinese Medicines Database (TCMD).⁹⁴ For each compound, a 3D structure with explicit stereochemistry defined by the 3D coordinates (each stereoisomer receives its own uniquely numbered structure in the case of ambiguous or racemic stereochemistry) as well as several identifiers and molecular descriptors are included in the downloadable files, either as individual SD files for single compounds or as an SD or CSV file containing the entire database. This database can be searched according to several different identifiers, chemical structure, and natural source.

Databases Focused on Traditional Chinese Medicines.

TCM Database@Taiwan. The Traditional Chinese Medicine Database@Taiwan (TCM Database@Taiwan) is the largest freely available source of traditional Chinese medicine (TCM) ingredient data.⁴⁴ This database contains over 60k TCM compounds from over 450 herb, mineral, and animal product TCMS compiled from a literature search including Chinese medical texts and dictionaries. The database is partitioned into 22 TCM usage classes (e.g., dampness-resolving medicinal and astringent medicinal) in addition to, in some cases, further subclasses based on traditional Chinese theories.⁴⁴ Searching the database by TCM category and TCM is possible on the Web site, as is an advanced search based on molecular properties and chemical structure. This database provides comprehensive ingredient-to-TCM mapping combined with 3D structures of each ingredient, including references to the original research articles for each compound and TCM. The

chemical structures of each compound are available with stereochemistry defined by the atom coordinates.

TCMID. The Traditional Chinese Medicine Integrated Database (TCMID)⁴⁶ stands out among the freely available NP databases because of its incorporation of data on drugs, targets, and diseases. This information is linked to NPs and the TCM herbs or formulas they are found in. In this way, this database focuses on the interface between traditional Chinese medicine and modern western medicine. The associations between the six components in the database (TCM formula, herb, compound, disease, drug, and target) are easily searchable on the Web site as well as available for download. In addition, other information such as chemical identifiers, the chemical structure including stereochemistry, and usage information can also be obtained via the TCMID. Visualization of these relationships is provided by the network display tool on the TCMID Web site that shows the ingredient–targets network, the ingredient–targets–drug–disease network, and the herb–target–disease network. For example, a visualization of the herb–target–disease network shows at a glance which ingredients are present in the herb, the targets of those compounds, and the diseases related to those targets. The information provided by the TCMID is thereby directly useful for drug discovery research, and can in particular be used to examine potential multitarget effects and molecular mechanisms.⁹⁵ The data on herbal ingredients come from the TCM Database@Taiwan, the Traditional Chinese Medicine Information Database (TCM-ID),⁹⁶ and the Encyclopedia of Traditional Chinese Medicines.⁹⁷ The linking of this information goes beyond that of the TCM Database@Taiwan in that the targets, drugs, and diseases are also included. The data for these three aspects come from DrugBank⁹⁸ and OMIM.⁹⁹

Chem-TCM. The Chemical Database of Traditional Chinese Medicine (Chem-TCM) is a commercial database focused on traditional Chinese medicine.⁴⁸ Like the TCMID, the Chem-TCM seeks to link traditional Chinese medicine to molecular targets of western medicine. These connections come from predicted activity based on models for each target or disease. For each compound, the Chem-TCM provides a calculated affinity for each of 28 major TCM categories and predicted activity against 41 therapeutic targets in western medicine. Chemical and botanical information is also provided. This database contains >12k compounds from around 350 herbs and allows structure and substructure search, text search, and creation of customized subsets of the database. Compound structures are available in SD format and stereochemistry information is provided.

HIT. The Herbal Ingredients' Targets database (HIT) connects active ingredients from herbs to their biological targets.⁵⁰ The Web site, which is freely available for academic use, provides an interface that links herbs to their ingredients and each ingredient to its target(s). The HIT database is searchable by keyword, including compound, herb, and protein target keywords, as well as by compound similarity and target similarity (based on protein sequence). The database is cross-linked to various relevant databases such as DrugBank, PDB, Therapeutic Targets Database (TTD),¹⁰⁰ Uniprot,¹⁰¹ and TCM-ID. The molecular structures can be downloaded from ZINC,^{91,102} including stereochemistry defined by chiral flags.

HIM. The Herbal Ingredients in vivo Metabolism database (HIM) has a distinct focus on ADMET data, in particular metabolism.⁵² Information about metabolism, bioactivity, and other ADMET properties was collected from primary and

secondary sources for all herbal ingredients present in the database. These data are cross-linked to relevant databases, including PubChem, HIT, and TCM-ID, and are freely available for academic use. Like the HIT, the HIM database can not be downloaded from the Web site but the molecular structures are available via ZINC. Stereochemistry is defined by the chiral flags in the 3D structures downloadable from ZINC. The contents of HIM are searchable on the Web site, either by text search, substructure search, or similarity search. A metabolism scheme is presented for each herbal active ingredient, including multiple generations of metabolites. The bioactivity data for each active ingredient may include a general classification, such as anticancer or antibacterial, rather than a specific protein target.

Databases and Libraries of African Natural Products.

AfroDb. Focused on NPs from African medicinal plants, the AfroDb is a relatively small NP library with large structural diversity.⁵⁴ This library, including 3D structures with stereochemistry (defined via chiral flags in the SD file) is freely available for download, either from the Supporting Information of the original publication (ref 54) or via the ZINC catalog of NPs.

AfroCancer. The African Anticancer Natural Products Library, AfroCancer,⁵⁶ contains experimentally confirmed anticancer, cytotoxic, and antiproliferative compounds from African medicinal plants. This database is freely downloadable from the Supporting Information of the original publication (ref 56) and includes 3D structures with stereochemistry defined by chirality flags. The pan-African natural products library (p-ANAPL)¹⁰³ provides information about the availability of the compounds in the AfroCancer database upon request.

AfroMalariaDB. The African Antimalarial Natural Products Library, AfroMalariaDB,⁵⁸ contains antimalarial NPs from 131 African plant species. The antiplasmodial and antimalarial activity of the compounds contained in this database has been measured in vitro and/or in vivo. In addition to activity, calculated physicochemical properties are provided. This database can be downloaded from the Supporting Information of the original publication (ref 58), including 3D structures with stereochemistry flags. Information about the availability of the compounds is provided by p-ANAPL upon request.

SANCDDB. SANCDDB, the South African natural compound database, contains NPs from South African plants and marine life. The first African database of NPs with a Web interface,⁶⁰ the SANCDDB is freely available and provides references, 3D structures, and other details for NPs compiled manually from the literature. Stereochemistry information is provided by isomeric SMILES. This database can be searched online by name, chemical structure, source organism, structural classification of the compound, physicochemical properties, and more. Structures can be downloaded in several formats from the search results, and users can contribute their own compounds to the database.

NANPDB. The Northern African Natural Products Database, NANPDB,⁶² is a freely accessible database containing NPs from Northern Africa. These NPs and their source organisms, biological activities and activity type (e.g., anticancer, antimalarial) were collected from the literature. The compounds come mostly from plants, but some stem from endophytes or animals, fungi, or bacteria. The contents of the NANPDB can be downloaded as a single file in either SMILES or SD format. Stereochemistry is included via chirality flags in

Table 2. Physical Natural Product Libraries

Supplier	(Sub-) set name	No. of molecules ^a	Chemistry-aware Web interface ^b	Composition	Online presence
Analyticon Discovery	MEGx – Purified natural products of microbial and plant origin	>4200	no	NP-only	110
Analyticon Discovery	NATx – Semisynthetic natural product-derived compounds	>26k	no	NPs and (semi-) synthetic compounds	110
Analyticon Discovery	FRGx – Fragments from nature	>200	no	NPs and (semi-) synthetic compounds	110
Analyticon Discovery	MACROx – Next generation macrocycles	>1800	no	NPs and (semi-) synthetic compounds	110
Ambinter and Greenpharma	Natural products	>8000	yes	NP-only	111,112
Ambinter and Greenpharma	Natural product derivatives	>11k	yes	(Semi-) synthetic compounds	111,112
InterBioScreen	Natural Compound (NC) Collection	>1300 natural compounds and >64k derivatives and analogs	no	NPs and (semi-) synthetic compounds; distinguishable by tags	113
InterBioScreen	Building Blocks	>13k	no	NPs and (semi-) synthetic compounds	113
InterBioScreen	Natural Scaffold Libraries	>500	no	NPs and (semi-) synthetic compounds	113
Developmental Therapeutic Program (DTP), NCI/NIH	Natural Products Set IV	>400	no	NP-only	114
TimTec	Natural Product Library (NPL)	~800	yes	NP-only	115
TimTec	Natural Derivatives Library (NDL)	~3000	yes	NPs and (semi-) synthetic compounds	115
TimTec	Flavonoids	~500	yes	NPs and (semi-) synthetic compounds	115
TimTec	ExtendedDB Flavonoid Derivatives	>4000	yes	NPs and (semi-) synthetic compounds	115
TimTec	Gossypol Derivatives	~100	yes	NPs and (semi-) synthetic compounds	115
Pi Chemicals	Natural Products Catalog	~2400 (~1900 natural compounds)	no	NPs and semisynthetic compounds; distinguishable by tags	116
p-ANAPL Library		>500	no	NP-only	
Selleck Chemicals	Natural Products	~130	no	NP-only	117
TargetMol	Natural Compound Library	~850	no	NP-only	118
AK Scientific	Natural Products	~250	yes	NP-only	119
AK Scientific	Synthetic-Additives	~130	yes	NPs and (semi-) synthetic compounds	119
MicroSource Discovery Systems	Natural Products Collection (NatProd)	~800	no	NPs and (semi-) synthetic compounds	120
Specs	Natural Products	~750	yes	NPs and (semi-) synthetic compounds	121
Sequoia Research Products		>2300	no	NPs and (semi-) synthetic compounds	122
Labseeker	Natural Compounds	>5300	yes	NPs and (semi-) synthetic compounds	123
Pharmeks	Screening Compounds	>340k (>2600 natural compounds and derivatives)	yes	NPs and (semi-) synthetic compounds; distinguishable by tags	124
Pharmeks	Building Blocks	>12k	yes	NPs and (semi-) synthetic compounds	124
Princeton BioMolecular Research	Macrocycles	>1500	yes	NPs and (semi-) synthetic compounds	125
Biopurify Phytochemicals	TCM Compounds Library	>2000	no	NPs and (semi-) synthetic compounds	126
INDOFINE Chemical Company	Natural Products, Flavonoids, Coumarins, etc.	~1900	no	NPs and (semi-) synthetic compounds	127
MedChem Express	Natural Product Library	>200	no	NPs and (semi-) synthetic compounds	128

^aNumber of molecules reported on the Web site of the database provider or supplied in the original files. ^bIndicates whether a chemistry-aware Web interface and search functionality (such as exact structure, substructure and similarity search) is provided.

the SD files and isomeric SMILES. The Web interface includes search functionality and allows submission of data to be included in the NANPDB. In addition to a basic search by compound name, organism or keyword, the NANPDB Web site offers a structure search based on either substructure or similarity.

Other Focused and Smaller-Sized Databases. *NPACT*. Cancer is the focus of NPACT,⁶⁴ the Naturally occurring Plant-based Anticancer Compound–Activity–Target database. This database is unusual in that it provides bioactivities of the NPs against over 300 cancer cell lines and protein targets. NPACT describes approximately 5200 compound–cell line and approximately 2000 compound–target interactions.⁶⁴ Stereo-

chemistry is defined by chiral flags in the downloadable MOL files, and the database is searchable, including a similarity search for structures. The entries are supplemented with references to other databases, including HIT and PubChem, as well as structures, properties and bioactivity data.

NPCARE. The focus of the freely available Database of Natural Products for CAncer gene REgulation (NPCARE) is on the relationship between NPs and cancer.⁶⁶ Specifically, this database contains information on gene expression and inhibition of cancer cells upon application of NPs. The database entries contain information on the cancer type, the NP and its source, the activity, the cancer cell line used, and the target gene or protein. Structural information is provided for more than 6500 compounds,⁶⁶ with stereochemistry defined by chiral flags in the SD files. Structures of more than 1500 of these compounds are available for bulk download, although without stereochemistry information.

TIPdb. The Taiwan indigenous plant database (TIPdb) contains anticancer, antituberculosis, and antiplatelet phytochemicals from plants indigenous to Taiwan.⁶⁸ This freely available database includes 3D structures, which are referred to as the TIPdb-3D.⁶⁹ The online database is searchable by plant and chemical names, database ID, and activity. Bioactivity information is available upon searching or browsing the TIPdb; however, this information is not present in the downloadable SD files.

Natural Products in PubChem Substance Database. The PubChem database contains NPs and their associated bioactivity data. This information can be accessed from the PubChem Substance database by the following query: "MLSMR[*SRC*] AND NP[*CMT*]".⁷¹ The resulting set of NPs consists of around 3000 unique chemical structures.^{104,105} Bioactivity data against 666 protein targets and other molecular targets is available for most of these NPs. Stereochemistry is defined via chiral flags in the downloadable structures.

StreptomeDB. The StreptomeDB is focused on NPs produced by streptomycetes,⁷³ with data assembled from the literature, the Novel Antibiotics Data Base,¹⁰⁶ and KNAP-SAcK.^{107,108} Stereochemistry is defined by chiral flags in the SD file. The StreptomeDB is freely available online and is searchable by name, structure, substructure, structural similarity and scaffold, as well as other aspects such as compound properties and activity. Phylogenetic classification of streptomycetes species is also included and is browsable via the Web site.

UEFS Natural Products. The Natural Products database of The State University of Feriera De Santana (UEFS) in Brazil contains structures of around 500 NPs collected from the literature. This data set is included in ZINC and does not have its own Web site.

NuBBE Database. Focused on NPs and derivatives from plants and microorganisms native to Brazil, the NuBBE database is a freely accessible online database.⁷⁵ This database provides, in addition to chemical compound information, pharmacological and toxicological data. The search capabilities include structure, compound identity, NMR shifts, physicochemical properties, and biological source. All structures can be downloaded in bulk as 3D structures in MOL2 file format.

Carotenoids Database. The Carotenoids Database⁷⁷ concentrates on natural carotenoids, currently containing over 1100 of this class of compound from nearly 700 source organisms. The data contained in this online database were obtained from the literature and include chemical information,

source organisms, and biological function of the compounds. Structures, including stereochemical information encoded in the isomeric SMILES and the InChI, are provided but can only be downloaded one molecule at a time.

AntiBase. The AntiBase database⁷⁹ is a comprehensive compilation of over 40k NPs, primarily with antimicrobial activity. The data provided in AntiBase come from the primary and secondary literature and include physicochemical properties, spectroscopic data, biological data such as activity and toxicity, and the biological source. Stereochemistry information is included. This database is commercially available in several software formats. The search capabilities of AntiBase include NMR shift search in addition to structure and text search.

DMNP. The Dictionary of Marine Natural Products (DMNP)¹⁰⁹ is a subset of the DNP that contains around 55k marine NPs and their derivatives. This database is available both online and as a combination of book and CD-ROM.

MarinLit. The MarinLit database⁸² is a comprehensive collection of marine NPs from the literature. The record for each literature article is comprised of bibliographic information, keywords, taxonomy of marine organisms, any compounds published for the first time in that article, and collection location of these new NPs. Compound records additionally include a full structure with stereochemistry, identifiers such as InChI, and chemical descriptors. Through the Web interface, the database can be queried on any of these aspects as well as on a combination of multiple parameters. In conjunction with its text and structure search capabilities, MarinLit provides additional search options for the purpose of dereplication.

■ PHYSICAL NATURAL PRODUCT LIBRARIES

Most vendors offer physical libraries that include a mix of NPs with synthetic and/or semisynthetic compounds. Catalogs consisting exclusively of natural products are rare. The molecular structures of purchasable natural products are generally provided free of charge and downloadable in bulk from the vendors' Web sites. An overview of physical natural product libraries is provided in Table 2. The overview is not comprehensive and is limited to catalogs that explicitly mention NPs and provide chemical structures or chemistry-aware search functionality on their Web sites.

Collections Consisting Entirely of Natural Products.

AnalytiCon Discovery. AnalytiCon Discovery provides a continuously growing collection of purified NPs. This collection, named MEGx, contains over 4200 highly pure compounds of known chemical structure. Many of the microbial compounds in this collection are exclusive to AnalytiCon, which isolates new NPs at a rate of around 500 novel compounds per year. In addition to the NP subset MEGx, AnalytiCon offers a semisynthetic NP-derived compound subset (NATx) with over 26k compounds, a macrocycles subset (MACROx) with over 1800 compounds, and a subset of fragments from nature (FRGx) with over 200 fragments.

Ambinter and Greenpharma. Ambinter and Greenpharma offer a set of approximately 8000 natural compounds. These include a number of different phytochemical families, alkaloids being the most well-represented thereof due to their vast structural diversity. In addition, these companies also offer a set of approximately 11k NP derivatives.

InterBioScreen. The Natural Compound (NC) collection of InterBioScreen consists of over 1300 NPs. In addition, this collection includes more than 64k NP derivatives and analogs.

These are annotated and hence can be distinguished from the genuine NPs. The majority of the compounds come from plant species, while 5 to 10% are of microbial origin and around 5% originate in marine life. Uncommon compounds are present in the NC collection as well, including allelopathic agents, unusual classes of phytoalexins, and specific sex attractants. Inter-BioScreen also offers a library of over 13k building blocks of natural and synthetic origin as well as more than 500 natural scaffolds for compound synthesis.

Developmental Therapeutic Program (DTP) Natural Products, NCI/NIH. The Developmental Therapeutic Program (DTP) of the National Cancer Institute (NCI) of the National Institutes of Health (NIH) provides a collection of NPs (Natural Products Set IV) including an SD file of 2D structures. This set contains over 400 NPs selected from the nearly 140k compounds in the DTP Open Repository based on structural diversity, availability, origin, and purity.

TimTec. TimTec's Natural Product Library (NPL) contains 800 pure NPs, mainly of plant origin but also from animal, bacterial, and fungal sources. TimTec additionally provides a Natural Derivatives Library (NDL) of over 3000 compounds, including natural derivatives, seminatural compounds, and NP analogs. The compounds in this library were selected from the literature and in-house data. A further subset of around 500 flavonoid derivatives is based on nine flavonoid core structures, and the Extended Flavonoid Derivatives database contains over 4000 compounds. A small library of gossypol derivatives is available as well.

Other Vendors. PI Chemicals offers a library of approximately 1900 natural compounds (annotated) and 400 semisynthetic chemicals. The p-ANAPL is a physical library of over 500 NPs, mostly flavonoids, found in African medicinal plants. Selleck Chemicals provides a NP library containing around 130 NPs from plant, marine, and microbial sources. TargetMol's Natural Compound Library includes around 850 compounds from many sources, including microorganisms, plants, and animals. The NP subset from AK Scientific contains around 250 naturally derived compounds, and a separate catalog of synthetics and additives includes over 100 flavonoids, food additives/preservatives, and vitamins.

Mixed Collections of Natural Products and Natural Product Derivatives. Several vendors offer NPs as part of a larger collection of compounds. In these cases it is often difficult or not possible to separate the NPs from NP derivatives and synthetic compounds. The Natural Products Collection (NatProd) from MicroSource Discovery Systems, containing 800 NPs and derivatives, includes natural compounds from plant, animal, and microbiological sources. Specs provides a collection of approximately 750 NPs, either isolated or synthesized, and NP derivatives of varying complexity from marine, plant, and microbial sources. Larger collections are available from Sequoia Research Products, a company specializing in biochemicals and other NPs with over 2300 compounds in its database, and Labseeker, a product distribution platform for over 5300 natural, semisynthetic, and synthetic compounds. The compound collection of Pharmeks is a diverse, mostly heterocyclic collection of organic molecules. Of the over 340k screening compounds in this library, over 2600 are natural compounds and derivatives thereof. In addition, Pharmeks offers over 12k building blocks, both natural and synthetic.

There are some collections with specialized focus as well. Princeton BioMolecular Research offers libraries of macrocyclic

molecules, with a total of over 1500 such compounds. These compounds are a mixture of NPs, semisynthetic NP derivatives, and completely synthetic molecules. TCM compounds are available from Biopurify Phytochemicals' TCM Compounds Library, which contains over 2000 natural compounds. INDOFINE Chemical Company has a focus on flavonoids, offering around 1900 NPs and semisynthetic compounds that include flavonoids, flavones, isoflavones, flavanones, coumarins, chromones, chalcones, and lipids. In addition, MedChem Express offers around 200 bioactive NPs for which preclinical research and clinical trials have indicated bioactivity and safety.

■ COVERAGE AND REACH OF CHEMICAL STRUCTURES DEPOSITED IN NATURAL PRODUCT LIBRARIES

The coverage, reach and overlap of molecular libraries was determined by counting the unique InChIs (without the stereochemistry and fixed hydrogen layers) among the sets of standardized molecules (counterions of salts removed and compounds neutralized with the Wash function in MOE).⁸⁵

Coverage of Free and Commercial Virtual Natural Product Libraries. The number of known NPs is around 250k, measured based on the freely available virtual natural product libraries as well as the DNP (i.e., all libraries indicated in Table 1 as being included in the analysis). More NPs can be found in freely accessible virtual NP databases (i.e., the subset of all free libraries indicated in Table 1) than are in the DNP (Figure 2a). The DNP contains about 53% of the compounds

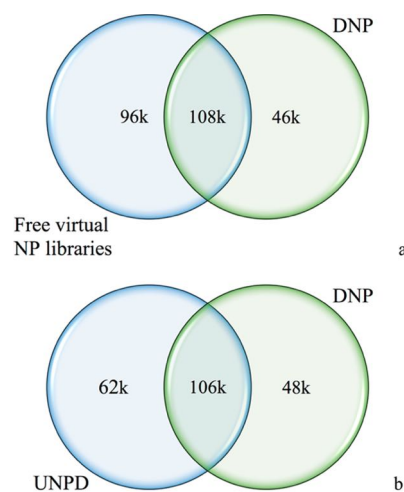


Figure 2. Overlap between the DNP and (a) the freely accessible virtual NP libraries or (b) the UNPD.

contained in the free libraries, whereas 70% of the compounds in the DNP can be found in at least one free library. This large overlap between the DNP and the freely accessible virtual NP databases stems primarily from the UNPD, the most comprehensive freely available, fully downloadable virtual NP database (Figure 2b). Although Super Natural II is larger than the UNPD, it was not used in the comparisons because it is not downloadable.

Coverage of molecular scaffolds appears to go along with database size. The DNP, UNPD and all freely accessible virtual NP databases together contain structures based on about 38k, 45k and 58k different Murcko scaffolds (calculated with RDKit¹²⁹ in KNIME¹³⁰), respectively. Flavonoids, steroids,

anthraquinones, coumarins and indoles are among the most populated scaffolds of all of these data sets. It is important to note that despite the good coverage of molecular structures by free databases, there are other advantages to commercial databases such as richer annotation.

Number of Readily Purchasable Natural Products and Derivatives. One consideration for virtual screening of NPs is what portion of the compounds present in any of the virtual NP libraries (“known NPs”) would be readily purchasable for experimental testing (“readily purchasable NPs”). The number of NPs offered by NP-only physical libraries is fairly low (around 11k unique compounds). However, many more NPs are available from vendors of mixed physical libraries (i.e., libraries containing NPs and (semi-) synthetic compounds). A comprehensive resource representing the purchasable chemical space is the ZINC database.^{91,102} The ZINC subset of readily purchasable compounds (downloadable as the ZINC “in-stock” subset) consists of around 7.3 M compounds. The overlap of this ZINC subset with the known NPs indicates that approximately 25k NPs (10% of all known NPs) covering more than 5600 different Murcko scaffolds are readily purchasable (Figure 3) from at least one of more than 100

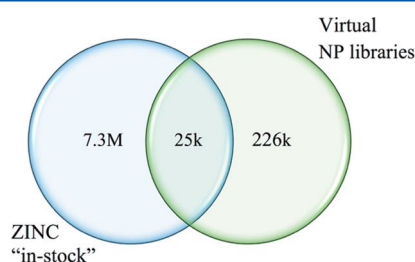


Figure 3. Comparison of the content of virtual NP libraries and the ZINC “in-stock” subset.

vendors identified as NP suppliers (Table 3). Nine of the largest of these suppliers each offer over 5000 readily purchasable NPs. In comparison, a recent study on the purchasable chemical space indicated that 36% of all NPs included in the TCM Database@Taiwan, the DNP and the StreptomeDB are purchasable if one is prepared to also use on-demand sourcing, extraction and synthesis services, which involve longer lead times and higher costs.¹³¹

When small deviations in molecular structure are allowed (i.e., Tanimoto coefficient based on ECFP4-like Morgan2 fingerprints equal to 0.7 or higher; calculated with RDKit), the coverage of virtual NP libraries by the ZINC “in-stock” subset increases (Figure 4) to 24% (~58k). These compounds are likely to be NP derivatives or analogs.

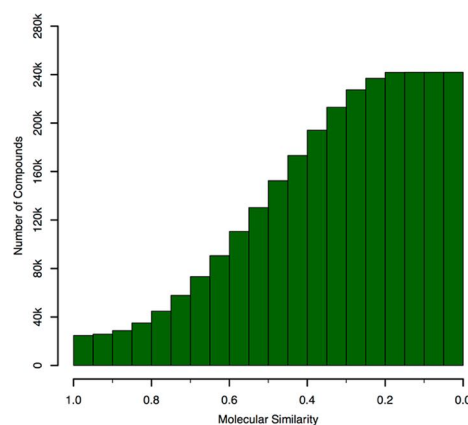


Figure 4. Cumulative histogram of maximum molecular similarity (Tanimoto coefficient) for the compounds in virtual NP libraries compared to the ZINC “in-stock” subset. The bars in the histogram represent the number of known NPs with a maximum molecular similarity greater than or equal to the bin threshold.

Comparing the molecular weight (MW) and logP of readily purchasable vs known NPs (Figure 5) indicates that readily purchasable NPs have a lower median molecular weight (267 vs 424 Da) and logP (2.18 vs 2.92). Here readily purchasable is defined as being contained in the overlap between virtual NP libraries and the ZINC “in-stock” subset. The size of available NPs is relevant because small, fragment-sized NPs can be used as starting points for fragment-based drug discovery. Out of the known NPs, about 23% (~57k of ~250k) are fragment-sized (MW less than 300 Da). Moreover, fragment-sized compounds make up 57% (~14k) of readily purchasable NPs.

Another category of compounds of particular interest for drug discovery are macrocycles. Because of their constrained conformations, macrocycles can provide an entropic binding advantage. The benefit of macrocycles extends to NPs, and in

Table 3. Numbers of Natural Products Readily Purchasable from Suppliers.^a

Number of readily purchasable natural products	Suppliers
>5000	Molport, TimTec, AK Scientific, Tetrahedron Scientific, BOC Sciences, FineTech Industry, Sigma-Aldrich, Specs, National Cancer Institute (NCI)
3000–5000	Fluorochem, Nanjing Kaimubo Pharmatech Company, Hong Kong Chemere, Oxchem Corporation, BePharm, Zelinsky Institute, Combi-Blocks, Debye Scientific, Matrix Scientific, WuXi AppTec, Ark Pharm, Bide Pharmatech, BioSynth, InterBioScreen, Labseeker, StruChem, Alfa-Aesar
2000–3000	AstaTech, Enamine, Oakwood Chemical, Frontier Scientific Services, Alfa Chemistry, Key Organics, Apollo Scientific, W&J PharmaChem, AnalytiCon Discovery, Acros Organics, Pi Chemicals, Syntharise Chemical
1000–2000	Toronto Research Chemicals, Capot Chemical, Rostar, INDOFINE Chemical Company, Alinda, Pharmeks, Innovapharm, Syntho-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
100–1000	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 NPs and the compounds present from a particular vendor in the “in-stock” subset of ZINC.

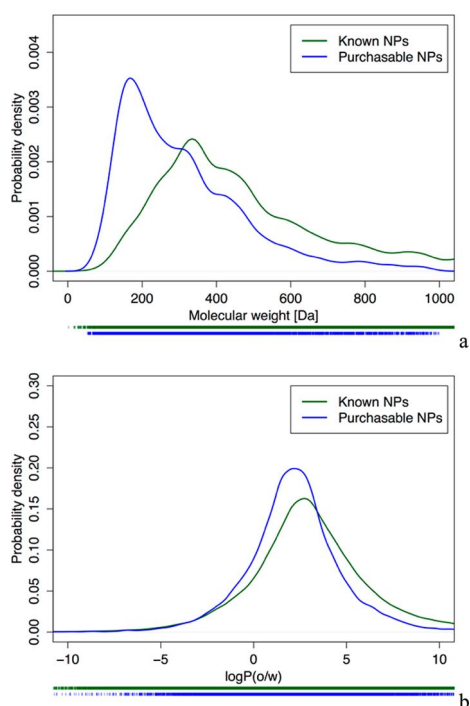


Figure 5. Distributions of (a) MW and (b) logP for known NPs (represented by the free virtual NP libraries and the DNP) vs readily purchasable NPs (overlap between known NPs and the ZINC “in-stock” subset).

fact a number of existing macrocyclic drugs are NPs or NP derivatives. Macrolides, which are derived from NPs, make up a large portion of the macrocyclic drugs on the market, including the majority of the orally administered macrocycles.¹³² In addition, the TCM Database@Taiwan has been shown to be highly enriched for a number of macrocyclic structures not present in large chemical databases such as ChEMBL, PubChem, and SciFinder.¹³³ Around 14% (~35k) and 13% (~33k) of known NPs contain a ring with >7 and >11 atoms, respectively. Fewer of the readily purchasable NPs are macrocycles; approximately 800 and 700 have a ring containing >7 and >11 atoms, respectively.

The ChEMBL^{134,135} is a widely used source of compound information, in particular biological data, for over 1.7 million compounds. Around 16% (~40k) of known NPs can be found in the current version of the ChEMBL. Though NPs make up only a small portion of the ChEMBL, this coverage of known NPs indicates that this freely available database is a good resource for information on the bioactivity of NPs.

CONCLUSIONS

Over 250k known NPs can be found in virtual NP libraries. Though there are many smaller virtual libraries of NPs with different areas of focus, there are larger, more comprehensive databases as well, both commercial and freely available. We found the overlap between the commercial and freely available virtual libraries in terms of NP structures to be quite large. In addition, a number of libraries of purchasable NPs provide corresponding structures that can also be used for virtual screening. Of the approximately 250k known NPs, i.e. those available in virtual NP libraries, around 25k are readily purchasable. In addition, NP derivatives and analogs are readily

purchasable for a further 10k to 30k known NPs, estimated based on molecular similarity.

The use of virtual NP libraries contributes to effective computer-guided drug discovery, enabling more efficient use of NPs as starting points for drug development. The many NP libraries that are already available are valuable resources, and many of these continue to grow in size.

AUTHOR INFORMATION

Corresponding Author

*J. Kirchmair. E-mail: kirchmair@zbh.uni-hamburg.de. Tel.: +49 (0)40 42838 7303.

ORCID

Ya Chen: 0000-0001-5273-1815

Christina de Bruyn Kops: 0000-0001-8890-2137

Johannes Kirchmair: 0000-0003-2667-5877

Author Contributions

#These authors contributed equally to this work.

Funding

Y.C. was supported by the China Scholarship Council (201606010345).

Notes

The authors declare no competing financial interest.

ABBREVIATIONS

ADME, absorption, distribution, metabolism, excretion; HTS, high-throughput screening; MW, molecular weight; NP, natural product; TCM, traditional Chinese medicine; QSAR, quantitative structure–activity relationship

REFERENCES

- (1) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79*, 629–661.
- (2) Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J. The Re-Emergence of Natural Products for Drug Discovery in the Genomics Era. *Nat. Rev. Drug Discovery* **2015**, *14*, 111–129.
- (3) Atanasov, A. G.; Waltenberger, B.; Pferschy-Wenzig, E.-M.; Linder, T.; Wawrosch, C.; Uhrin, P.; Temml, V.; Wang, L.; Schwaiger, S.; Heiss, E. H.; Rollinger, J. M.; Schuster, D.; Breuss, J. M.; Bochkov, V.; Mihovilovic, M. D.; Kopp, B.; Bauer, R.; Dirsch, V. M.; Stuppner, H. Discovery and Resupply of Pharmacologically Active Plant-Derived Natural Products: A Review. *Biotechnol. Adv.* **2015**, *33*, 1582–1614.
- (4) White, N. J. Qinghaosu (Artemisinin): The Price of Success. *Science* **2008**, *320*, 330–334.
- (5) Su, X.-Z.; Miller, L. H. The Discovery of Artemisinin and the Nobel Prize in Physiology or Medicine. *Sci. China: Life Sci.* **2015**, *58*, 1175–1179.
- (6) Roemer, T.; Krysan, D. J. Antifungal Drug Development: Challenges, Unmet Clinical Needs, and New Approaches. *Cold Spring Harbor Perspect. Med.* **2014**, *4*, a019703.
- (7) ROCS, Version 3.2.1.4; OpenEye Scientific Software: Santa Fe, NM, 2015. <http://www.eyesopen.com>.
- (8) Chackalamannil, S.; Wang, Y.; Greenlee, W. J.; Hu, Z.; Xia, Y.; Ahn, H.-S.; Boykow, G.; Hsieh, Y.; Palamanda, J.; Agans-Fantuzzi, J.; Kurowski, S.; Graziano, M.; Chintala, M. Discovery of a Novel, Orally Active Himbacine-Based Thrombin Receptor Antagonist (SCH 530348) with Potent Antiplatelet Activity. *J. Med. Chem.* **2008**, *51*, 3061–3064.
- (9) Mottamal, M.; Zheng, S.; Huang, T. L.; Wang, G. Histone Deacetylase Inhibitors in Clinical Studies as Templates for New Anticancer Agents. *Molecules* **2015**, *20*, 3898–3941.
- (10) Choi, C.-I. Sodium-Glucose Cotransporter 2 (SGLT2) Inhibitors from Natural Products: Discovery of Next-Generation Antihyperglycemic Agents. *Molecules* **2016**, *21*, 1136.

- (11) Henrich, C. J.; Beutler, J. A. Matching the Power of High Throughput Screening to the Chemical Diversity of Natural Products. *Nat. Prod. Rep.* **2013**, *30*, 1284–1298.
- (12) Stratton, C. F.; Newman, D. J.; Tan, D. S. Cheminformatic Comparison of Approved Drugs from Natural Product versus Synthetic Origins. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 4802–4807.
- (13) Camp, D.; Garavelas, A.; Campitelli, M. Analysis of Physicochemical Properties for Drugs of Natural Origin. *J. Nat. Prod.* **2015**, *78*, 1370–1382.
- (14) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzler, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting Biologically Relevant Chemical Space: A Structural Classification of Natural Products (SCONP). *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17272–17277.
- (15) Cragg, G. M.; Newman, D. J. Biodiversity: A Continuing Source of Novel Drug Leads. *Pure Appl. Chem.* **2005**, *77*, 7–24.
- (16) Prachayasittikul, V.; Worachartcheewan, A.; Shoombuatong, W.; Songtawe, N.; Simeon, S.; Prachayasittikul, V.; Nantasenamat, C. Computer-Aided Drug Design of Bioactive Natural Products. *Curr. Top. Med. Chem.* **2015**, *15*, 1780–1800.
- (17) Rodrigues, T.; Reker, D.; Schneider, P.; Schneider, G. Counting on Natural Products for Drug Design. *Nat. Chem.* **2016**, *8*, 531–541.
- (18) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S. Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.
- (19) Hert, J.; Irwin, J. J.; Lagner, C.; Keiser, M. J.; Shoichet, B. K. Quantifying Biogenic Bias in Screening Libraries. *Nat. Chem. Biol.* **2009**, *5*, 479–483.
- (20) Sukuru, S. C. K.; Jenkins, J. L.; Beckwith, R. E. J.; Scheiber, J.; Bender, A.; Mikhailov, D.; Davies, J. W.; Glick, M. Plate-Based Diversity Selection Based on Empirical HTS Data to Enhance the Number of Hits and Their Chemical Diversity. *J. Biomol. Screening* **2009**, *14*, 690–699.
- (21) van Hattum, H.; Waldmann, H. Biology-Oriented Synthesis: Harnessing the Power of Evolution. *J. Am. Chem. Soc.* **2014**, *136*, 11853–11859.
- (22) Morrison, K. C.; Hergenrother, P. J. Natural Products as Starting Points for the Synthesis of Complex and Diverse Compounds. *Nat. Prod. Rep.* **2014**, *31*, 6–14.
- (23) Ma, D.-L.; Chan, D. S.-H.; Leung, C.-H. Molecular Docking for Virtual Screening of Natural Product Databases. *Chem. Sci.* **2011**, *2*, 1656–1665.
- (24) Feher, M.; Schmidt, J. M. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.
- (25) Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small Molecules of Different Origins Have Distinct Distributions of Structural Complexity That Correlate with Protein-Binding Profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18787–18792.
- (26) Lee, M. L.; Schneider, G. Scaffold Architecture and Pharmacophoric Properties of Natural Products and Trade Drugs: Application in the Design of Natural Product-Based Combinatorial Libraries. *J. Comb. Chem.* **2001**, *3*, 284–289.
- (27) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F. Statistical Investigation into the Structural Complementarity of Natural Products and Synthetic Compounds. *Angew. Chem., Int. Ed.* **1999**, *38*, 643–647.
- (28) Jayaseelan, K. V.; Moreno, P.; Truszkowski, A.; Ertl, P.; Steinbeck, C. Natural Product-Likeness Score Revisited: An Open-Source, Open-Data Implementation. *BMC Bioinf.* **2012**, *13*, 106.
- (29) Ikram, N. K. K.; Durrant, J. D.; Muchtaridi, M.; Zaaludin, A. S.; Purwitasari, N.; Mohamed, N.; Rahim, A. S. A.; Lam, C. K.; Normi, Y. M.; Rahman, N. A.; Amaro, R. E.; Wahab, H. A. A Virtual Screening Approach for Identifying Plants with Anti HSN1 Neuramidase Activity. *J. Chem. Inf. Model.* **2015**, *55*, 308–316.
- (30) Liu, Y.; Huang, L.; Ye, H.; Lv, X. 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.* **2016**, *27*, 939–948.
- (31) Tietjen, I.; Ntie-Kang, F.; Mwananzi, P.; Onguéné, P. A.; Scull, M. A.; Idowu, T. O.; Ogundaini, A. O.; Meva'a, L. M.; Abegaz, B. M.; Rice, C. M.; Andrae-Marobela, K.; Brockman, M. A.; Brumme, Z. L.; Fedida, D. Screening of the Pan-African Natural Product Library Identifies Ixoratanin A-2 and Boldine as Novel HIV-1 Inhibitors. *PLoS One* **2015**, *10*, e0121099.
- (32) Mohamed, A.; Nguyen, C. H.; Mamitsuka, H. Current Status and Prospects of Computational Resources for Natural Product Dereplication: A Review. *Briefings Bioinf.* **2016**, *17*, 309–321.
- (33) Füllbeck, M.; Michalsky, E.; Dunkel, M.; Preissner, R. Natural Products: Sources and Databases. *Nat. Prod. Rep.* **2006**, *23*, 347–356.
- (34) Tung, C.-W. Public Databases of Plant Natural Products for Computational Drug Discovery. *Curr. Comput.-Aided Drug Des.* **2015**, *10*, 191–196.
- (35) Lagunin, A. A.; Goel, R. K.; Gawande, D. Y.; Pahwa, P.; Glorizova, T. A.; Dmitriev, A. V.; Ivanov, S. M.; Rudik, A. V.; Konova, V. I.; Pogodin, P. V.; Druzhilovsky, D. S.; Poroikov, V. V. Chemo- and Bioinformatics Resources for in Silico Drug Discovery from Medicinal Plants beyond Their Traditional Use: A Critical Review. *Nat. Prod. Rep.* **2014**, *31*, 1585–1611.
- (36) Blunt, J.; Munro, M.; Upjohn, M. The Role of Databases in Marine Natural Products Research. In *Handbook of Marine Natural Products*; Fattorusso, E., Gerwick, W. H., Tagliatalata-Scafati, O., Eds.; Springer Netherlands: Dordrecht, 2012; pp 389–421.
- (37) Blunt, J. W.; Munro, M. H. G. Is There an Ideal Database for Natural Products Research? In *Natural Products*; Osbourn, A., Goss, R. J., Carter, G. T., Eds.; John Wiley & Sons, Inc., 2014; pp 413–431.
- (38) Dictionary of Natural Products (DNP). <http://dnp.chemnetbase.com> (accessed Apr 7, 2017).
- (39) Reaxys; Elsevier: New York, <https://www.reaxys.com> (accessed Jul 17, 2017).
- (40) Banerjee, P.; Erehman, J.; Gohlke, B.-O.; Wilhelm, T.; Preissner, R.; Dunkel, M. Super Natural II – a Database of Natural Products. *Nucleic Acids Res.* **2015**, *43*, D935–D939.
- (41) Super Natural II. http://bioinf-applied.charite.de/supernatural_new (accessed Apr 10, 2017).
- (42) Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PLoS One* **2013**, *8*, e62839.
- (43) Universal Natural Products Database (UNPD). <http://pkuxj.pku.edu.cn/UNPD> (accessed Oct 17, 2016).
- (44) Chen, C. Y.-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening In Silico. *PLoS One* **2011**, *6*, e15939.
- (45) TCM Database@Taiwan. <http://tcm.cmu.edu.tw> (accessed Oct 17, 2016).
- (46) Xue, R.; Fang, Z.; Zhang, M.; Yi, Z.; Wen, C.; Shi, T. TCMID: Traditional Chinese Medicine Integrative Database for Herb Molecular Mechanism Analysis. *Nucleic Acids Res.* **2013**, *41*, D1089–D1095.
- (47) Traditional Chinese Medicine Integrated Database (TCMID). www.megabionet.org/tcmid (accessed Oct 19, 2016).
- (48) Ehrman, T. M.; Barlow, D. J.; Hylands, P. J. Phytochemical Informatics of Traditional Chinese Medicine and Therapeutic Relevance. *J. Chem. Inf. Model.* **2007**, *47*, 2316–2334.
- (49) Chem-TCM. www.chemtcm.com (accessed Apr 10, 2017).
- (50) Ye, H.; Ye, L.; Kang, H.; Zhang, D.; Tao, L.; Tang, K.; Liu, X.; Zhu, R.; Liu, Q.; Chen, Y. Z.; Li, Y.; Cao, Z. HIT: Linking Herbal Active Ingredients to Targets. *Nucleic Acids Res.* **2011**, *39*, D1055–D1059.
- (51) Herbal Ingredients' Targets database (HIT). <http://lifecenter.sgst.cn/hit> (accessed Apr 13, 2017).
- (52) Kang, H.; Tang, K.; Liu, Q.; Sun, Y.; Huang, Q.; Zhu, R.; Gao, J.; Zhang, D.; Huang, C.; Cao, Z. HIM-Herbal Ingredients in-Vivo Metabolism Database. *J. Cheminf.* **2013**, *5*, 28.
- (53) Herbal Ingredients in-vivo Metabolism database (HIM). <http://binfo.shmtu.edu.cn:8080/him> (accessed Apr 13, 2017).

- (54) Ntie-Kang, F.; Zofou, D.; Babiaka, S. B.; Meudom, R.; Scharfe, M.; Lifongo, L. L.; Mbah, J. A.; Mbaze, L. M.; Sippl, W.; Efang, S. M. AfroDb: A Select Highly Potent and Diverse Natural Product Library from African Medicinal Plants. *PLoS One* **2013**, *8*, e78085.
- (55) AfroDb. <http://african-compounds.org/about/afrodb/> (accessed Oct 18, 2016).
- (56) Ntie-Kang, F.; Nwodo, J. N.; Ibezim, A.; Simoben, C. V.; Karaman, B.; Ngwa, V. F.; Sippl, W.; Adikwu, M. U.; Mbaze, L. M. Molecular Modeling of Potential Anticancer Agents from African Medicinal Plants. *J. Chem. Inf. Model.* **2014**, *54*, 2433–2450.
- (57) AfroCancer. <http://african-compounds.org/about/afrocancer/> (accessed Feb 10, 2017).
- (58) Onguéné, P. A.; Ntie-Kang, F.; Mbah, J. A.; Lifongo, L. L.; Ndom, J. C.; Sippl, W.; Mbaze, L. M. 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.* **2014**, *4*, 6.
- (59) AfroMalariaDB. <http://african-compounds.org/about/afromalariaadb/> (accessed Feb 10, 2017).
- (60) Hatherley, R.; Brown, D. K.; Musyoka, T. M.; Penkler, D. L.; Faya, N.; Lobb, K. A.; Tastan Bishop, Ö. SANCDB: A South African Natural Compound Database. *J. Cheminf.* **2015**, *7*, 29.
- (61) South African Natural Compound Database (SANCDB). <http://sancdb.rubi.ru.ac.za> (accessed Feb 8, 2017).
- (62) Ntie-Kang, F.; Telukunta, K. K.; Döring, K.; Simoben, C. V.; Moumbock, A. F. A.; Malange, Y. I.; Njume, L. E.; Yong, J. N.; Sippl, W.; Günther, S. NANPDB: A Resource for Natural Products from Northern African Sources. *J. Nat. Prod.* **2017**, *80*, 2067–2076.
- (63) Northern African Natural Products Database (NANPDB). www.african-compounds.org/nanpdb (accessed Apr 5, 2017).
- (64) Mangal, M.; Sagar, P.; Singh, H.; Raghava, G. P. S.; Agarwal, S. M. NPACT: Naturally Occurring Plant-Based Anti-Cancer Compound-Activity-Target Database. *Nucleic Acids Res.* **2013**, *41*, D1124–D1129.
- (65) Naturally Occurring Plant-based Anticancerous Compound-Activity-Target Database (NPACT). <http://crdd.osdd.net/raghava/npact> (accessed Apr 13, 2017).
- (66) Choi, H.; Cho, S. Y.; Pak, H. J.; Kim, Y.; Choi, J.-Y.; Lee, Y. J.; Gong, B. H.; Kang, Y. S.; Han, T.; Choi, G.; Cho, Y.; Lee, S.; Ryoo, D.; Park, H. NPCARE: Database of Natural Products and Fractional Extracts for Cancer Regulation. *J. Cheminf.* **2017**, *9*, 2.
- (67) Database of Natural Products for Cancer Gene Regulation (NPCARE). <http://silver.sejong.ac.kr/npicare> (accessed Feb 20, 2017).
- (68) Lin, Y.-C.; Wang, C.-C.; Chen, I.-S.; Jheng, J.-L.; Li, J.-H.; Tung, C.-W. TIPdb: A Database of Anticancer, Antiplatelet, and Anti-tuberculosis Phytochemicals from Indigenous Plants in Taiwan. *Sci. World J.* **2013**, *2013*, 736386.
- (69) Tung, C.-W.; Lin, Y.-C.; Chang, H.-S.; Wang, C.-C.; Chen, I.-S.; Jheng, J.-L.; Li, J.-H. TIPdb-3D: The Three-Dimensional Structure Database of Phytochemicals from Taiwan Indigenous Plants. *Database* **2014**, *2014*, bau055.
- (70) Taiwan Indigenous Plant Database (TIPdb). <http://cwtung.kmu.edu.tw/tipdb> (accessed Oct 19, 2016).
- (71) Hao, M.; Cheng, T.; Wang, Y.; Bryant, H. S. Web Search and Data Mining of Natural Products and Their Bioactivities in PubChem. *Sci. China: Chem.* **2013**, *56*, 1424–1435.
- (72) PubChem Substance. <http://ncbi.nlm.nih.gov/pcsubstance> (accessed Apr 7, 2017).
- (73) Klementz, D.; Döring, K.; Lucas, X.; Telukunta, K. K.; Erleben, A.; Deubel, D.; Erber, A.; Santillana, I.; Thomas, O. S.; Bechthold, A.; Günther, S. StreptomeDB 2.0—an Extended Resource of Natural Products Produced by Streptomycetes. *Nucleic Acids Res.* **2016**, *44*, D509–D514.
- (74) StreptomeDB. www.pharmaceutical-bioinformatics.de/streptomedb (accessed Apr 13, 2017).
- (75) Valli, M.; dos Santos, R. N.; Figueira, L. D.; Nakajima, C. H.; Castro-Gamboa, I.; Andricopulo, A. D.; Bolzani, V. S. Development of a Natural Products Database from the Biodiversity of Brazil. *J. Nat. Prod.* **2013**, *76*, 439–444.
- (76) Núcleo de Bioensaios, Biossíntese e Ecofisiologia de Produtos Naturais (NuBBE). <http://nubbe.iq.unesp.br/portal/nubbedb.html> (accessed Apr 19, 2017).
- (77) Yabuzaki, J. Carotenoids Database: Structures, Chemical Fingerprints and Distribution among Organisms. *Database* **2017**, *2017*, bax004.
- (78) Carotenoids Database. www.carotenoiddb.jp (accessed Apr 10, 2017).
- (79) Laatsch, H. *AntiBase: The Natural Compound Identifier*; Wiley-VCH: Weinheim, 2017.
- (80) AntiBase. <http://wwwuser.gwdg.de/~hlaatsc/antibase.htm> (accessed Apr 10, 2017).
- (81) Dictionary of Marine Natural Products. <http://dmnp.chemnetbase.com> (accessed Apr 10, 2017).
- (82) MarinLit. <http://pubs.rsc.org/marinlit> (accessed Apr 10, 2017).
- (83) *InChI*, version 1.05; IUPAC: Research Triangle Park, NC, 2017.
- (84) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. *InChI*, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7*, 23.
- (85) *Molecular Operating Environment (MOE)*, Version 2016.08; Chemical Computing Group ULC: Montreal, QC, 2016.
- (86) *Dictionary of Natural Products, Version 19.1 [CD-ROM]*; CRC Press: London, 2010.
- (87) Flemming, C. Searching for Natural Products in Reaxys. Presented at Reaxys Webinar [Online], Jan 27, 2015. <https://www.youtube.com/watch?v=vJKXsDDhRyk> (accessed Jul 17, 2017).
- (88) Kanehisa, M.; Goto, S.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. Data, Information, Knowledge and Principle: Back to Metabolism in KEGG. *Nucleic Acids Res.* **2014**, *42*, D199–D205.
- (89) Caspi, R.; Altman, T.; Billington, R.; Dreher, K.; Foerster, H.; Fulcher, C. A.; Holland, T. A.; Keseler, I. M.; Kothari, A.; Kubo, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. A.; Ong, Q.; Paley, S.; Subhraveti, P.; Weaver, D. S.; Weerasinghe, D.; Zhang, P.; Karp, P. D. The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc Collection of Pathway/Genome Databases. *Nucleic Acids Res.* **2014**, *42*, D459–D471.
- (90) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorn Dahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Res.* **2013**, *41*, D801–D807.
- (91) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
- (92) Shen, J.; Xu, X.; Cheng, F.; Liu, H.; Luo, X.; Shen, J.; Chen, K.; Zhao, W.; Shen, X.; Jiang, H. Virtual Screening on Natural Products for Discovering Active Compounds and Target Information. *Curr. Med. Chem.* **2003**, *10*, 2327–2342.
- (93) Qiao, X.; Hou, T.; Zhang, W.; Guo, S.; Xu, X. A 3D Structure Database of Components from Chinese Traditional Medicinal Herbs. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 481–489.
- (94) He, M.; Yan, X.; Zhou, J.; Xie, G. Traditional Chinese Medicine Database and Application on the Web. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 273–277.
- (95) Fang, Z.; Lu, B.; Liu, M.; Zhang, M.; Yi, Z.; Wen, C.; Shi, T. Evaluating the Pharmacological Mechanism of Chinese Medicine Si-Wu-Tang through Multi-Level Data Integration. *PLoS One* **2013**, *8*, e72334.
- (96) Wang, J.; Zhou, H.; Han, L.; Chen, X.; Chen, Y.; Cao, Z. Traditional Chinese Medicine Information Database. *Clin. Pharmacol. Ther.* **2005**, *78*, 92–93.
- (97) Zhou, J.; Xie, G.; Yan, X. *Encyclopedia of Traditional Chinese Medicines - Molecular Structures, Pharmacological Activities, Natural Sources and Applications*; Springer: Berlin, 2011; Vol. 6.
- (98) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive

Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672.

(99) Online Mendelian Inheritance in Man (OMIM) database. www.ncbi.nlm.nih.gov/omim (accessed Apr 10, 2017).

(100) Zhu, F.; Han, B.; Kumar, P.; Liu, X.; Ma, X.; Wei, X.; Huang, L.; Guo, Y.; Han, L.; Zheng, C.; Chen, Y. Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* **2010**, *38*, D787–D791.

(101) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.

(102) ZINC15. <http://zinc15.docking.org> (accessed May 26, 2017).

(103) Ntie-Kang, F.; Amoa Onguéné, P.; Fotso, G. W.; Andrae-Marobela, K.; Bezabih, M.; Ndom, J. C.; Ngadjui, B. T.; Ogundaini, A. O.; Abegaz, B. M.; Meva'a, L. M. Virtualizing the P-ANAPL Library: A Step towards Drug Discovery from African Medicinal Plants. *PLoS One* **2014**, *9*, e90655.

(104) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202–D1213.

(105) PubChem Compound – NCBI. <https://www.ncbi.nlm.nih.gov/pccompound> (accessed Mar 23, 2017).

(106) Novel Antibiotics Data Base. <http://www.antibiotics.or.jp/journal/database/database-top.htm> (accessed Apr 7, 2017).

(107) Nakamura, Y.; Afendi, F. M.; Parvin, A. K.; Ono, N.; Tanaka, K.; Hirai Morita, A.; Sato, T.; Sugiura, T.; Altaf-Ul-Amin, M.; Kanaya, S. KNAPSAcK Metabolite Activity Database for Retrieving the Relationships between Metabolites and Biological Activities. *Plant Cell Physiol.* **2014**, *55*, e7.

(108) Afendi, F. M.; Okada, T.; Yamazaki, M.; Hirai-Morita, A.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-Ul-Amin, M.; Darusman, L. K.; Saito, K.; Kanaya, S. KNAPSAcK Family Databases: Integrated Metabolite-Plant Species Databases for Multifaceted Plant Research. *Plant Cell Physiol.* **2012**, *53*, e1.

(109) *Dictionary of Marine Natural Products with CD-ROM*; Blunt, J. W.; Munro, M. H. G., Eds.; Chapman and Hall/CRC: Boca Raton, FL, 2007.

(110) AnalytiCon Discovery. www.ac-discovery.com (accessed Apr 21, 2017).

(111) Ambinter. www.ambinter.com (accessed Jun 2, 2017).

(112) GreenPharma. www.greenpharma.com (accessed Jun 2, 2017).

(113) InterBioScreen. www.ibscreen.com (accessed Apr 21, 2017).

(114) Natural Products Set IV of the Developmental Therapeutic Program (DTP), NCI/NIH. http://dtp.cancer.gov/organization/dscb/obtaining/available_plates.htm (accessed Oct 20, 2016).

(115) TimTec. www.timtec.net (accessed Oct 19, 2016).

(116) PI Chemicals. www.pipharm.com (accessed May 5, 2017).

(117) Selleck Chemicals. www.selleckchem.com (accessed Feb 13, 2017).

(118) TargetMol. www.targetmol.com (accessed May 17, 2017).

(119) AK Scientific. www.aksci.com (accessed Apr 19, 2017).

(120) MicroSource Discovery Systems. www.msdiscovery.com (accessed Oct 24, 2016).

(121) Specs. www.specs.net (accessed Mar 30, 2017).

(122) Sequoia Research Products. www.seqchem.com (accessed Oct 20, 2016).

(123) LabSeeker. www.labseeker.com (accessed Apr 19, 2017).

(124) Pharmeks. www.pharmeks.com (accessed Apr 18, 2017).

(125) Princeton BioMolecular Research. www.princetonbio.com (accessed Feb 3, 2017).

(126) Biopurify Phytochemicals. www.biopurify.com (accessed Apr 1, 2017).

(127) INDOFINE Chemical Company. www.indofinechemical.com (accessed Apr 24, 2017).

(128) Medchem Express. www.medchemexpress.com (accessed May 18, 2017).

(129) RDKit: Open-Source Cheminformatics, Version 2016.03.1, 2016; <http://www.rdkit.org/>.

(130) Berthold, M. R.; Cebren, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The

Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization*; Springer: Berlin, 2007; pp 319–326.

(131) Lucas, X.; Grüning, B. A.; Bleher, S.; Günther, S. The Purchasable Chemical Space: A Detailed Picture. *J. Chem. Inf. Model.* **2015**, *55*, 915–924.

(132) Giordanetto, F.; Kihlberg, J. Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? *J. Med. Chem.* **2014**, *57*, 278–295.

(133) Kramer, C.; Podewitz, M.; Ertl, P.; Liedl, K. R. Unique Macrocycles in the Taiwan Traditional Chinese Medicine Database. *Planta Med.* **2015**, *81*, 459–466.

(134) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.

(135) ChEMBL version 23. <https://www.ebi.ac.uk/chembl> (accessed Jun 6, 2017).

Our 2017 review on NP databases was well-received by the scientific community and we were invited to publish an updated and extended analysis as a book chapter in a volume of *Progress in the Chemistry of Organic Natural Products* dedicated to cheminformatics in NP research (D3).

The book chapter includes several additional virtual NP databases and uses a more fine-grained categorization into encyclopedic and/or general NP databases, databases focused on traditional medicines, databases focused on a specific habitat or geographical region, and databases focused on specific organisms, biological activities or specific NP class.

Importantly, the book chapter also includes information on physicochemical properties of NPs from the different virtual and physical NP databases. This information originates primarily from our study presented in [Chapter 3.1 \(Characterization of Physicochemical and Structural Properties of Natural Products, D4\)](#).

[D3] Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Resources for Chemical, Biological, and Structural Data on Natural Products. In *Progress in the Chemistry of Organic Natural Products*; Kinghorn, A. D., Falk, H., Gibbons, S., Kobayashi, J., Asakawa, Y., Liu, J.-K., Eds.; Springer, 2019; Vol. 110, pp 37–71.

Available at https://doi.org/10.1007/978-3-030-14632-0_2.

Y. Chen, C. de Bruyn Kops and J. Kirchmair conceptualized the work. Y. Chen analyzed the literature and collected, curated and analyzed all the data. Y. Chen wrote the largest part of the manuscript. J. Kirchmair supervised this work.

Reprinted by permission from Springer Nature:
Springer, Cham. *Progress in the Chemistry of Organic Natural Products 110* by Kinghorn, A. D., Falk, H., Gibbons, S., Kobayashi, J., Asakawa, Y., Liu, J.-K. (eds). Copyright Springer Nature Switzerland AG 2019

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 (SANCDB)	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
SANCDB	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 antiplasmodial 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
BIOFACQJIM	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 (SANCDDB)

The SANCDDB [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.

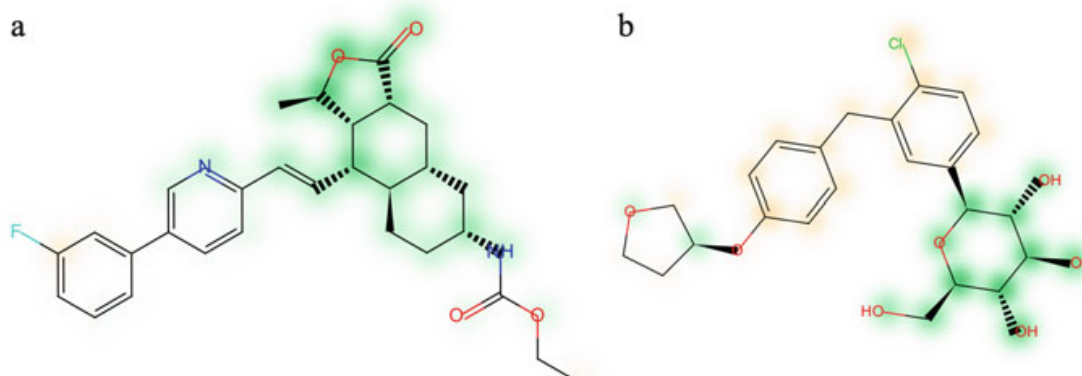


Fig. 1 Similarity maps of (a) vorapaxar and (b) empagliflozin. Green-highlighted atoms contribute to the classification of a molecule as a natural product; orange-highlighted atoms contribute to the classification of a molecule as a synthetic compound. Adapted from [78] (CC BY 4.0; <https://creativecommons.org/licenses/by/4.0>)

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.

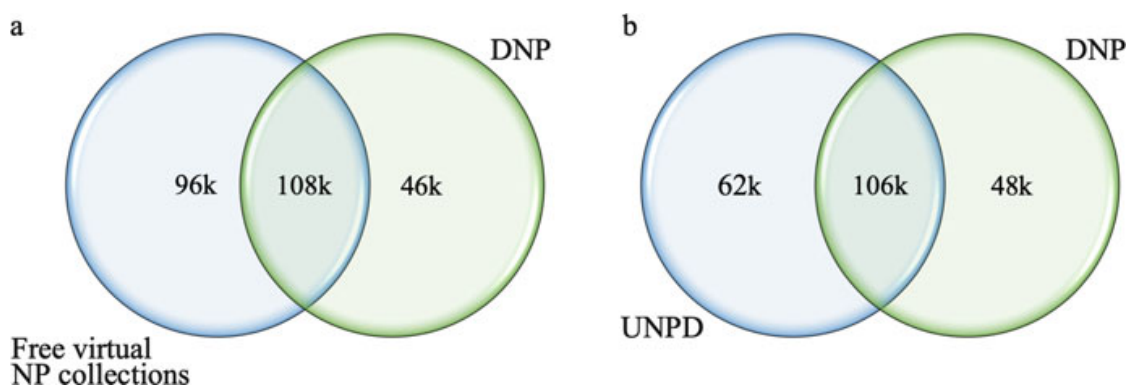


Fig. 2 The overlap between the Dictionary of Natural Products (DNP) and (a) the freely accessible virtual natural product collections or (b) the Universal Natural Products Database (UNPD). Reprinted with permission from [10]. Copyright 2017 American Chemical Society

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

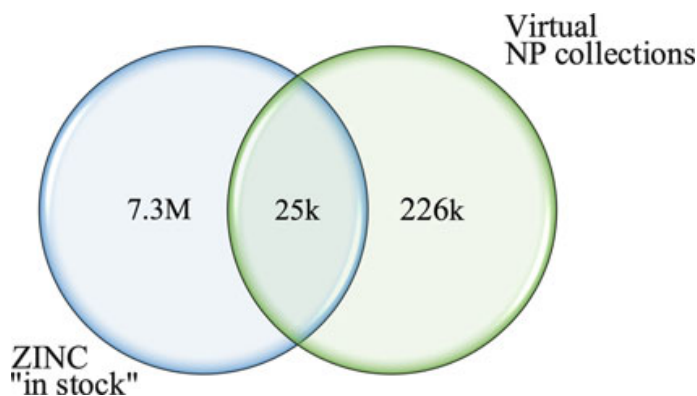


Table 3 Numbers of natural products readily purchasable from suppliers^a

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

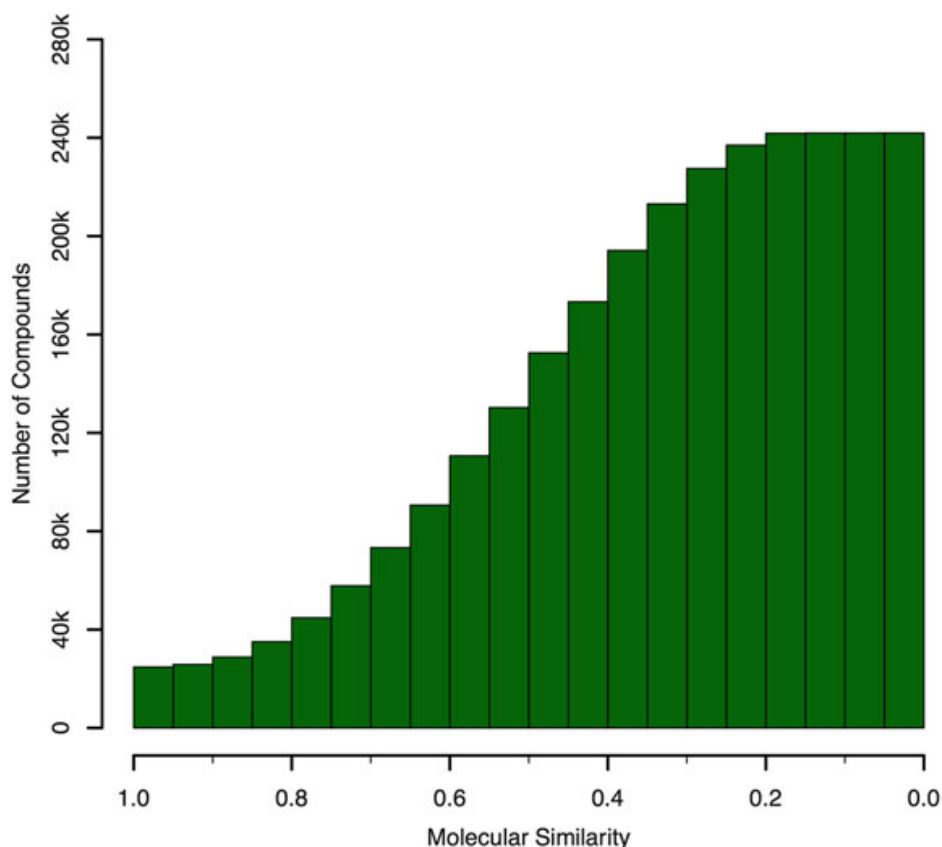


Fig. 4 Cumulative histogram of maximum molecular similarity (Tanimoto coefficient) for the compounds in virtual natural product libraries compared to the ZINC “in-stock” subset. The bars in the histogram represent the number of known natural products with a maximum molecular similarity greater than or 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/impapat>
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, Efange 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, Tastan 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 Streptomyces. *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>

2.2. Chemical Data Preprocessing and Molecular Descriptor Calculation

One compound may be represented by different molecular structures. For this reason it is essential to develop and employ a robust protocol for the preprocessing (including structure standardization) of molecular structures prior to conducting any analyses or model building.

This subsection describes a generalized approach to preprocessing molecular structures. Further information is provided in the individual Methods sections of the publications presented as part of the [Results](#) section of this dissertation.

Today, a plethora of tools for molecular structure preprocessing are at our disposal. The work of this thesis builds primarily on the use of the KNIME data analytics platform [15], MOE [16] and RDKit [17] (via Python scripts).

The general data preprocessing workflow involves the:

- conversion of different input formats to Simplified Molecular Input Line Entry Specification (SMILES) [18]
- addition of hydrogens
- removal of salt components
- removal of compounds with uncommon elements (usually, any compounds consisting of elements other than H, B, C, N, O, F, Si, P, S, Cl, Br and I)
- removal of compounds with molecular weight above or below a defined threshold
- assignment of appropriate formal charges (in general, neutralization)
- assignment of a standard tautomer state or enumeration of possible tautomers
- generation of the IUPAC International Chemical Identifier (InChIs) [19,20] or canonicalization of SMILES notations (ignoring or considering stereochemical information as appropriate)

Following data preprocessing, molecular descriptors or fingerprints are calculated. More specifically, for studying the physicochemical properties of NPs and comparing them with those of approved drugs ([Chapter 3.1, D4](#)) we employed MOE and RDKit for descriptor calculation (controlled via KNIME). For our analysis of a shape-based approach for predicting the macromolecular targets of complex small molecules ([Chapter 3.3, D6](#)) we used RDKit for descriptor calculation (controlled via Python scripts).

2.3. Chemical Space Analysis

One of the most frequently employed methods for the characterization and comparison of the chemical space covered by individual data sets is principal component analysis (PCA), which is explained in more detail in [D1](#). Briefly, PCA is a data projection method which reduces high-dimensional data into a low-dimensional space (usually a two-dimensional (2D) or 3D space for the purpose of visualization). The loadings of the principal components (PCs) indicate the correlation between a descriptor and a PC, thus helping to understand whether

certain descriptors are positively or negatively correlated with PCs and have strong effect on PCs. When two datasets are different on one PC, the main descriptors responsible for this difference can be identified.

The PCAs reported in this thesis are based on simple, physically meaningful molecular descriptors such as molecular weight, $\log P$, topological polar surface area, number of hydrogen bond acceptors, number of hydrogen bond donors, number of heavy atoms, fraction of rotatable bonds, number of nitrogen atoms, number of oxygen atoms, number of acidic atoms, number of basic atoms, sum of formal charges, number of aromatic atoms, number of chiral centers, and number of rings. The list of molecular descriptors used for each PCA study can be found in [D4](#) and [D5](#).

2.4. Rule-Based Approaches

For the characterization of NP databases, we elaborated and employed several rule-based approaches, described in more detail in [Chapter 3.1 \(D4\)](#). Because sugar or sugar-like moieties are in most cases not of interest in drug discovery, we developed “SugarBuster” ([D4](#)), a rule based method for deglycosylation. The rules are SMILES arbitrary target specification (SMARTS) [[21](#)] patterns that are designed to remove sugar and sugar-like moieties. The rules are defined as follows:

- five-membered aliphatic ring moieties with
 - exactly one heteroatom in the ring AND
 - all carbons forming the ring being a member of only one ring AND
 - at least two substituents with EITHER
 - ◆ two oxygen atoms attached directly to the ring OR
 - ◆ one oxygen atom and one nitrogen atom attached directly to the ring
- six-membered aliphatic ring moieties with
 - a maximum of one heteroatom in the ring AND
 - all carbons forming the ring being a member of only one ring AND
 - at least three substituents with EITHER
 - ◆ three oxygen atoms attached directly to the ring OR
 - ◆ two oxygen atoms and one nitrogen atom attached directly to the ring.

The algorithm removes the defined sugar and sugar-like fragments and returns the largest aglycon component (fragment with the highest number of heavy atoms) of the input molecule.

For classifying NPs into different NP classes, substructure matching by SMARTS patterns was also applied. The NP classes that can be identified include:

- basic alkaloids, defined as any NP with at least one basic nitrogen atom
- extended definition of alkaloids, defined as any NP with at least one basic nitrogen atom, quaternary nitrogen atom or carboxamide moiety
- phenols
- phenols or phenol ethers
- steroids, defined by the core ring system

- flavonoids, as well as several different subclasses of flavonoids namely anthocyanidins, chalcones, flavandiols, flavanols, flavanones, flavanonols, flavones, flavonols, and isoflavones, based on the scaffolds defined in Figure 26.35 of ref 22. The structures and the developed SMARTS patterns of the flavonoid subclasses are shown in Table 1.

2.5. Machine Learning Methods

Machine learning is becoming one of the most common in silico methods for drug discovery. Although several different machine learning algorithms have been explored during these studies (including support vector machines, random forest, and artificial neural networks), we converged to using a random forest algorithm (implemented in the scikit-learn library [23,24]) for modeling as it consistently obtained the best performance.

In the development of NP-Scout (Chapter 3.2, D5), large sets of NPs and synthetic molecules were merged and then randomly split into a training set and a test set with a ratio of 4:1 using the “train_test_split” method in the “model_selection” module of scikit-learn. In fingerprint space, structurally distinct molecules may have identical fingerprints. For this reason, deduplication, based on fingerprints, was separately performed for all NPs and all synthetic molecules in the training data. Any fingerprints present in both the NP and synthetic molecules subsets were removed in order to avoid conflicting class labels.

For the training set, 10-fold cross-validation was used to train models and stratified K-Folds cross-validation (“StratifiedKFold”) was performed to keep the same distribution of classes for each fold. Cross-validated grid-search was employed to select the best hyperparameters to train the models, while in this study (D5) hyperparameter tuning had no obvious improvement of the performance.

Random forest classifiers were generated using default settings, except for “n_estimators”, which was set to “100”, and “class_weight”, which was set to “balanced”. Three different models were built based on three different descriptors sets, namely Morgan2 fingerprints (1024 bits) [25,26] and MACCS keys (166 bits) calculated with RDKit, and 206 2D physicochemical property descriptors calculated with MOE.

2. Data Resources and Methods

Table 1. Examples of different subclasses of flavonoids and their substructure SMARTS patterns.

Subclasses of flavonoids	Structures from ref 22	SMARTS patterns
Anthocyanidins		<chem>[cR1]1([OX2&!R])[cR1][cR1]([OX2&!R])[cR1]c2[oR1;+][cR1]([cR1]3[cR1][cR1][cR1]([OX2&!R])[cR1][cR1]3)[cR1]([OX2&!R])[cR1]c12</chem>
Chalcones		<chem>[cR1]1([OX2&!R])[cR1][cR1]([OX2&!R])[cR1][cR1]([OX2&!R)[cR1]1[#6&!R](=O)[#6&!R]=[#6&!R][cR1]1[cR1][cR1][cR1]([OX2&!R])[cR1][cR1]1</chem>
Flavandiols		<chem>[cR1]1([OX2&!R])[cR1][cR1]([OX2&!R])[cR1]c2[OR1][CR1]([cR1]3[cR1][cR1][cR1]([OX2&!R])[cR1][cR1]3)[CR1]([OX2&!R])[CR1]([OX2&!R])c12</chem>
Flavanols		<chem>[cR1]1([OX2&!R])[cR1][cR1]([OX2&!R])[cR1]c2[OR1][CR1]([cR1]3[cR1][cR1][cR1]([OX2&!R])[cR1][cR1]3)[CR1]([OX2&!R])[CR1]([*;!O])c12</chem>
Flavanones		<chem>[cR1]1([OX2&!R])[cR1][cR1]([OX2&!R])[cR1]c2[OR1][CR1]([cR1]3[cR1][cR1][cR1]([OX2&!R])[cR1][cR1]3)[CR1]([*;!O])[CR1](=O)c12</chem>
Flavanonols		<chem>[cR1]1([OX2&!R])[cR1][cR1]([OX2&!R])[cR1]c2[OR1][CR1]([cR1]3[cR1][cR1][cR1]([OX2&!R])[cR1][cR1]3)[CR1]([OX2&!R])[CR1](=O)c12</chem>
Flavones		<chem>[cR1]1([OX2&!R])[cR1][cR1]([OX2&!R])[cR1]c2[oR1][cR1]([cR1]3[cR1][cR1][cR1]([OX2&!R])[cR1][cR1]3)[cR1]([*;!O])[cR1](=O)c12</chem>
Flavonols		<chem>[cR1]1([OX2&!R])[cR1][cR1]([OX2&!R])[cR1]c2[oR1][cR1]([cR1]3[cR1][cR1][cR1]([OX2&!R])[cR1][cR1]3)[cR1]([OX2&!R])[cR1](=O)c12</chem>
Isoflavones		<chem>[cR1]1([OX2&!R])[cR1][cR1]([OX2&!R])[cR1]c2[oR1][cR1]([*;!O])[cR1]([cR1]3[cR1][cR1][cR1]([OX2&!R])[cR1][cR1]3)[cR1](=O)c12</chem>

In the validation step, the performance of the selected models was measured on the test set. Model performance was characterized utilizing the Matthews correlation coefficient (MCC) [27] and area under the receiver operating characteristic curve (AUC). The MCC is one of the most robust measures for evaluating the performance of binary classifiers, as it considers the proportion of all classes in the confusion matrix (i.e., true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)):

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$$

MCC values range from -1 to 1, where 1 indicates perfect prediction, -1 a total opposite prediction, and 0 a performance equal to random prediction. The AUC was used to measure how well the models are able to rank NPs early in a list. The AUC can range from 0 to 1, where 1 indicates a perfect model, 0 indicates the model reversed the classes, and 0.5 means the model has no class separation capacity.

The model's ability to identify NPs was also tested on the Dictionary of Natural Products [28] as an external validation set.

2.6. Three-Dimensional Shape-Based Similarity Method

The usage of molecular similarity methods in drug discovery is based on the so-called "similarity principle", which states that compounds with similar chemical structures are likely to share similar properties, such as biological activities.

Most NPs are in low degree of structural similarity with more conventional, synthetic compounds, which account for the bulk of the measured activity data. Compared to 2D similarity-based approaches, more complex similarity-based methods that compare molecules based on their 3D molecular shape are designed to recognize such distant structural similarity. In Chapter 3.3 (D6), we systematically explored the capacity of ROCS [29,30], a leading shape-based screening engine that also takes into account chemical feature distributions, to identify the macromolecular targets of "complex" small molecules based on a knowledge base of "non-complex" compounds with measured bioactivity data.

Molecular similarity was quantified separately by each of four similarity metrics implemented in ROCS: ShapeTanimoto score, TanimotoCombo score, RefTverskyCombo score, and FitTverskyCombo score. As suggested by their names, these metrics are either based on the Tanimoto or the Tversky coefficient. The Tanimoto coefficient and Tversky coefficient quantify the similarity of two molecules, f and g , based on their self-volume overlaps (I_f and I_g) and the volume overlap between the two molecules ($O_{f,g}$):

$$Tanimoto_{f,g} = \frac{O_{f,g}}{I_f + I_g - O_{f,g}}$$

$$Tversky_{f,g} = \frac{O_{f,g}}{\alpha*I_f + \beta*I_g}$$

The Tversky coefficient can be asymmetric (depending on the *alpha* and *beta* parameters chosen; normally $\alpha + \beta = 1$), hence allowing the emphasis of either substructure or superstructure matching.

The ShapeTanimoto score only considers the fit of the molecule shapes for the volume overlap, whereas the three "combo" scores additionally take the type and distribution of chemical features (color) into account. The ShapeTanimoto score ranges from 0 to 1, with a value of 1 indicating a perfect fit of molecular shapes. The "combo" scores typically range from 0 to 2, with equal weights applied to the shape and color components.

The RefTverskyCombo score assigns an *alpha* value of 0.95 to the query as the main self-overlap term. The FitTverskyCombo score, on the contrary, assigns a *beta* value of 0.95 to the fit molecule. Note that the RefTverskyCombo and FitTverskyCombo scores can have values greater than 2 because the overlap of the two compounds can be larger than a molecule's self-overlap.

ROCS was run with factory settings with the following exceptions: both "-besthits" and "-maxhits" were set to "o" in order to cause ROCS to retain all results. The "-rankby" option was set to "ShapeTanimoto", "TanimotoCombo", "RefTverskyCombo", or "FitTverskyCombo" in order to have the results ranked by the four similarity metrics. For experiments using the ShapeTanimoto score, the "-shapeonly" function was enabled in order to cause ROCS to align molecules by taking only molecular shape into account (and not color).

3. Results

3.1. Characterization of Physicochemical and Structural Properties of Natural Products

NPs can differ substantially from synthetic molecules with regard to their physicochemical and structural properties, and these differences have been assessed in a large number of studies [31–33]. However, the understanding of the physicochemical properties and the chemical space of NPs from distinct resources, backgrounds and domains is limited. For this reason, based on the NP collections available to us (D2), we conducted a comprehensive cheminformatics analysis to obtain a detailed picture of the physicochemical and structural properties of different NP data sets (D4).

Overall, we analyzed 18 virtual NP libraries and nine physical NP libraries, from which we compiled data sets of all known NPs (from all the virtual databases), and the readily obtainable NPs (from all the physical libraries and the known NPs presented in ZINC database [34,35]). Additionally, a set of approved drugs from NPs or NP derivatives and the data set of NPs with high quality X-ray crystal structures in the Protein Data Bank (PDB) [36,37] were also included in the analysis.

The study provides detailed information on the number of unique structures, exclusive structures, scaffolds, structures containing sugar or sugar-like moieties, different NP classes, and different physicochemical properties etc. of each data resource, and also provides insights on the overall chemical space covered by the most relevant data resources.

As part of this study, an algorithm that can remove sugars and sugar moieties from NPs, as well as a rule-based approach that can identify different major classes of NPs, were developed.

The details of this work are reported in the following publication.

[D4] Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58* (8), 1518–1532.

Available at <https://doi.org/10.1021/acs.jcim.8b00302>.

Y. Chen and J. Kirchmair conceptualized the work. Y. Chen analyzed the literature and collected, curated and analyzed all the data. This involved, among many other tasks, the validation and use of the tool “SugarBuster”, which was developed by M. Garcia de Lomana and N.-O. Friedrich, and validated by Y. Chen. Y. Chen wrote the largest part of the manuscript. J. Kirchmair supervised this work.

Reprinted with permission from

Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58* (8), 1518–1532.

Copyright 2018 American Chemical Society.

The supplementary information of this article can be found in [Appendix A](#).

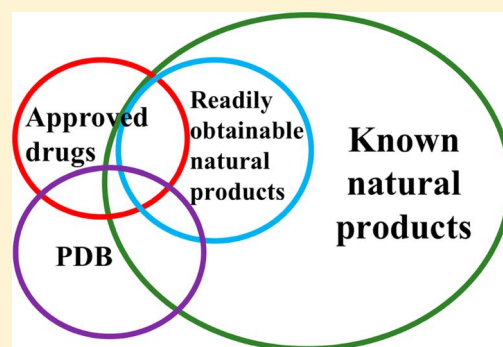
Characterization of the Chemical Space of Known and Readily Obtainable Natural Products

Ya Chen,¹ Marina Garcia de Lomana,¹ Nils-Ole Friedrich,¹ and Johannes Kirchmair^{*,1}

Center for Bioinformatics, Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany

S Supporting Information

ABSTRACT: Natural products remain one of the most productive sources of chemical inspiration for the development of new drugs. The structures of more than 250 000 natural products are available from public databases. At least 10% of these compounds are readily obtainable for experimental testing from commercial vendors and public research institutions. While the physicochemical properties of known natural products have been thoroughly studied and compared to those of drugs and other types of small molecules, the information available on the content, coverage, and relevance of individual virtual and physical natural product libraries is clearly limited. The aim of this study was the development of a detailed understanding of the coverage of chemical space by known and readily obtainable natural products and by individual natural product databases. For this purpose, we compiled comprehensive data sets of known and readily obtainable natural products from 18 virtual databases (including the Dictionary of Natural Products), nine physical libraries, and the Protein Data Bank (PDB). We also developed and employed an algorithm (“SugarBuster”) for the removal of sugars and sugar-like moieties, which are generally not in the focus of interest for drug discovery, from natural products. In addition, we devised a rule-based approach for the automated classification of natural products into natural product classes (alkaloids, steroids, flavonoids, etc.). Among the most important results of this study is the finding that the readily obtainable natural products are highly diverse and populate regions of chemical space that are of high relevance to drug discovery. In some cases, substantial differences in the coverage of natural product classes and chemical space by the individual databases are observed. More than 2000 natural products are identified for which at least one X-ray crystal structure of the compound in complex with a biomacromolecule is available from the PDB.



INTRODUCTION

Natural products (NPs) have a long history of use as components of traditional medicines. With the advent of synthetic organic chemistry in the early 19th century, a transition began from the use of complex mixtures of unknown content to the exploration and use of single bioactive natural products. Today, a large percentage of approved drugs are NPs or derived from NPs. According to a comprehensive analysis, 6% of all small-molecule drugs approved between 1981 and 2014 are unaltered NPs, 26% are NP derivatives, and 32% are NP mimetics and/or contain an NP pharmacophore.¹

In comparison to synthetic drug-like molecules, NPs stand out because of their enormous structural and physicochemical diversity.^{2–4} Rooted in their evolution-based specific biological purposes, NPs exhibit a wide range of biological activities in different organisms. Some NP scaffolds or substructures are therefore considered privileged structures.^{5,6}

A large number of NP libraries are available today. In a recent work, we analyzed the value of 25 virtual and 31 physical NP libraries for computer-guided drug discovery.⁷ The virtual NP libraries include encyclopedic databases such as the Dictionary of Natural Products (DNP)⁸ as well as many

specialized libraries focused, among others, on NPs related to traditional Chinese medicine (TCM) (e.g., TCM Database@Taiwan⁹), certain geographic regions (e.g., AfroDb¹⁰), or specific indications such as cancer (e.g., NPCARE¹¹). In total, these virtual NP libraries contain more than 250 000 unique NPs.⁷ Most of them can be used free of charge, for instance, for virtual screening.

A severe bottleneck in the research on NPs is the availability of material for testing. Only an estimated 10% of NPs that have their chemical structures deposited in virtual libraries are readily purchasable from commercial sources for timely experimental testing.⁷ There are also legal aspects to consider when sourcing NPs, in particular when transferring materials across national boundaries.¹² Here, computational methods that allow the identification of the most promising NPs for extraction, purification, (partial) synthesis, and biological testing are of high value to drug discovery. These methods in particular include virtual screening technologies for cherry-picking NPs with a high chance of exhibiting the desired

Received: May 18, 2018

Published: July 16, 2018

Table 1. Overview of Data Sets Investigated in This Work^a

Data set	Compounds ^{b,c}	Exclusive NPs ^d	Exclusive NPs [%] ^d	Murcko scaffolds ^b	Molecules with sugar-like moieties [%] ^{e,f}	Basic alkaloids [%] ^{g,h}	Alkaloids [%] ^{i,a}	Phenols [%] ^e	Phenols and phenol ethers [%] ^e	Steroids [%] ^e	Flavonoids [%] ^{e,i}	Scientific literature and/or online presence
Reference data sets												
Known NPs	208 166	ND	ND	50 366	20	16	26	27	37	8	2	
Readily obtainable NPs	25 524	ND	ND	5704	18	12	26	21	35	5	2	
Approved drugs	1867	ND	ND	1053	5	ND	ND	11	26	5	0	29,30
Newman and Cragg data set ^j	350	ND	ND	229	16	36	62	19	27	12	1	1
NPs of PDB	2060	ND	ND	510	ND	21	37	18	22	3	1	
Virtual NP databases: Encyclopedic and general databases												
DNP	128 757	36 364	28	28 557	17	11	20	24	34	7	2	33
UNPD	140 475	46 310	33	34 085	23	10	19	25	35	8	2	18,34
NPs of PubChem Substance Database	2760	1096	40	1031	5	12	52	20	50	2	2	35,36
Virtual NP databases related to TCM												
TCM Database@Taiwan	42 521	24 117	57	15 925	19	38	44	40	51	9	2	9,37
TCMID	10 073	688	7	3678	21	14	18	28	39	8	4	38,39
HIM	636	174	27	221	21	14	20	52	65	2	8	40,41
HIT	387	5	1	193	16	14	20	39	51	3	10	42,43
Virtual NP databases focused on specific geographical regions and bioactivities												
AfroDb	847	385	45	393	9	11	13	47	61	6	5	10, 44
AfroCancer	342	104	30	175	15	4	8	44	57	11	8	45,46
AfroMalariaDB	241	59	24	140	6	13	14	50	61	4	5	47,48
NANPDB	3297	602	18	1237	28	7	11	24	32	8	6	49,50
SANCDDB	534	52	10	281	16	17	20	31	40	14	2	51,52
TIPdb	6673	363	5	2339	25	9	14	37	51	6	8	53,54,55
NuBBE	1613	174	11	636	9	4	9	30	53	7	6	56,57
UEFS Natural Products	476	140	29	270	3	10	12	33	49	8	9	Via ZINC ^{27,28}
NPACT	1304	122	9	630	14	4	6	39	50	11	8	58,59
NPCARE	1436	103	7	802	13	8	17	30	41	9	6	11,60
StreptomeDB	3182	1014	32	1322	25	19	47	31	34	1	0	61,62
Physical NP libraries												
Ambinter and Greenpharma NPs	6058	4053	67	2331	19	12	21	34	48	8	5	63,64
AnalytiCon Discovery MEGx	3346	2445	73	1247	35	3	14	33	45	5	4	65
Pi Chemicals NPs	1244	193	16	561	32	14	20	36	48	10	7	66
InterBioScreen NPs	1067	48	4	485	20	25	35	18	37	15	2	67
TargetMol Natural Compound Library	674	114	17	319	27	16	23	31	42	11	5	68
p-ANAPL	443	296	67	191	15	5	6	58	72	5	11	69
Developmental Therapeutic Program (DTP), NCI/NIH NP Set IV	392	238	61	278	14	27	42	25	46	4	1	70
AK Scientific NPs	174	42	24	93	21	16	27	40	49	6	13	71
Selleck Chemicals NPs	155	8	5	92	20	15	23	41	52	5	12	72

^aAbbreviations: AfroCancer, the African Anticancer Natural Products Library; AfroDb, NPs from African medicinal plants; AfroMalariaDB, the African Antimalarial Natural Products Library; DNP, Dictionary of Natural Products; HIM, the Herbal Ingredients in Vivo Metabolism Database; HIT, the Herbal Ingredients' Targets Database; NANPDB, the Northern African Natural Products Database; NPACT, the Naturally Occurring Plant-Based Anticancer Compound-Activity-Target Database; NPCARE, Database of Natural Products for Cancer Gene Regulation; NuBBE, Nuclei of Bioassays, Ecophysiology and Biosynthesis of Natural Products Database; p-ANAPL, the Pan-African Natural Products Library; SANCDDB, the South African Natural Compound Database; StreptomeDB, NPs produced by streptomycetes; TCM Database@Taiwan, the Traditional Chinese Medicine Database@Taiwan; TCMID, the Traditional Chinese Medicine Integrated Database; TIPdb, the Taiwan Indigenous Plant Database; UEFS Natural Products, the natural products database of the State University of Feira De Santana; UNPD, the Universal Natural Products Database; ND, not determined. ^bNumber of unique molecular structures after data preprocessing (including the removal of sugars and sugar-like moieties with SugarBuster; see [Materials and Methods](#) for details). ^cThe numbers of compounds contained in the database source files are reported in ref 7. Differences between the numbers of compounds in the preprocessed data and the database source files are primarily related to the duplicate removal process. ^dNumber or percentage of NPs exclusive to a data resource within the categories "virtual NP databases" or "physical NP libraries". For example, 73% of all NPs contained in the AnalytiCon Discovery MEGx database are exclusively available via this provider. ^eThe color gradient indicates the percentages of unique compounds assigned to various NP classes and ranges from white (lowest percentage) to dark green (highest percentage). The color gradient should not be interpreted as indicative of "better" or "worse" values. NPs can be assigned to more than one NP class. ^fAny NP with at least one sugar or sugar-like moiety as defined by SugarBuster (see [Materials and Methods](#) for details). ^gAny NP with at least one basic nitrogen atom. ^hExtended definition of alkaloids: any NP with at least one basic nitrogen atom or quaternary nitrogen atom or carboxamide moiety. ⁱAny compound based on a characteristic scaffold as defined in Figure 26.35 of ref 73. ^jSubset of the Newman and Cragg data set containing all drugs approved between 1981 and 2014 that are NPs or NP derivatives.

biological activity, but also approaches for ADME, toxicity, and target prediction.^{6,13,14} However, cheminformatics methods

applied to drug discovery are generally developed on the basis of and for use with synthetic drug-like molecules.

In comparison with synthetic drug-like molecules, NPs can be of higher structural complexity, in particular with respect to stereochemical aspects and molecular shape (e.g., NPs tend to have a higher number of chiral centers^{15–17}). Therefore, computational methods may need to be modified to make them applicable to NPs. An interesting example of such a modification is an approach for predicting the mode of action of complex NPs by dissecting them into medium-sized fragments.¹³

A study by Ertl and Schuffenhauer⁴ comparing the physicochemical and structural properties of NPs with those of bioactive and organic drug-like molecules found that while there is a clear separation in the structural space that NPs and synthetic molecules populate, bioactive molecules were present in both groups. Importantly, an automated deglycosylation procedure was employed in this study to remove sugar moieties from NPs, as these primarily affect pharmacokinetic properties (they are generally metabolically labile) but are only rarely part of the pharmacophore and, as such, essential for bioactivity. Gu et al.¹⁸ compared the physicochemical property space of NPs (represented by the Universal Natural Products Database (UNPD)¹⁸) with that of FDA-approved drugs. They found that NPs are highly diverse in structure and that a large fraction of them obey the commonly applied definitions of drug-likeness. A study comparing Merck's NP collection, the company's sample collection, and the 200 top-selling drugs of 2006 found that NPs cover a much broader region of chemical space than compounds of the two other chemical origins.² Chen et al.¹⁹ carried out a comparative analysis of NPs, human metabolites, drugs, clinical candidates, and known bioactive compounds. They found that NPs and human metabolites cover different regions of chemical space compared with synthetic compounds. It was also found that NPs have the highest ring-system complexity among all of the investigated data sets and that NPs and human metabolites have more three-dimensional molecular shape diversity (related, e.g., to branching). El-Elimat et al.²⁰ compared the regions of chemical space covered by metabolites from fungi, cyanobacteria, and plants with that covered by FDA-approved anticancer drugs. They found a partial overlap of the regions of chemical space covered by metabolites from the three origins but also detected regions of chemical space populated only by compounds of a specific origin. One of their main conclusions was that the regions of chemical space covered by the investigated metabolites and FDA-approved anticancer drugs align well. Muigg et al.²¹ compared the regions of chemical space of NPs harvested from marine and terrestrial organisms with that of synthetic compounds. They found clear differences in the regions of chemical space covered by the compounds of these three origins. For example, NPs extracted from marine organisms tend to be large and highly flexible compared with synthetic compounds, which are generally smaller and less flexible. NPs originating from terrestrial organisms were found to be often large and rigid. Very recently, Shang et al.²² analyzed the chemical space covered by marine natural products. They found long chains and macrocyclic structures to be more prominent among marine natural products than terrestrial ones. Lucas et al.²³ compiled a data set of over 68 million unique purchasable compounds and overlaid them with a set of more than 227 000 known NPs. They found the NPs to have higher shape and stereogenic complexity than fragment-like compounds, drug-like compounds, and inhibitors

of protein–protein interactions. Some bioactive substructures common to NPs were rarely observed for marketed drugs.

The differences in the regions of chemical space covered by NPs and synthetic molecules of different origin have been thoroughly studied. However, the coverage of the chemical space relevant to drug discovery by readily obtainable NPs remains largely unknown, although it is highly relevant to early drug discovery and to computer-guided screening of NP libraries in particular. Few studies of the chemical space covered by individual NP libraries have been reported. For example, Yongye et al.²⁴ analyzed the scaffold diversity of five publicly available NP databases (three libraries of compound vendors, the NP subset of ZINC,²⁵ and TCM Database@Taiwan⁹), one combinatorial library set, and a general screening library from Maybridge.²⁶ They found that the investigated NP databases differ with respect to coverage of chemical space and scaffold diversity.

Most published studies do not include a mechanism for removing sugars and sugar-like moieties (which in the context of drug discovery are generally not in the focus of interest) from NPs. Also, most published analyses of the structural diversity of NPs and molecular scaffolds are centered around the counting of cyclic systems, while some biologically active scaffolds do not have a ring system or consist of a combination of ring systems and linkers. Here an analysis of the abundance of NP classes (alkaloids, flavonoids, steroids, etc.) would be of interest.

With this work, we aim to go beyond the range and thoroughness of previous studies to develop a comprehensive, detailed, and clean picture of the chemical space covered by NPs and its representation by molecular libraries of readily obtainable NPs. In order to do so, we have developed an algorithm for the automated removal of sugars and sugar-like moieties from molecular structures and devised a rule-based approach for the automated identification of major classes of NPs.

RESULTS

Data Sets. We collected and curated data from different sources to compile two comprehensive data sets, one representing the chemical space covered by known NPs and the other describing the chemical space covered by readily obtainable NPs. An overview of all of the data sets analyzed in this work is provided in Table 1. When interpreting this table, it is important to consider that the color gradient is not indicative of “better” or “worse” values and that NPs can be assigned to more than one NP class.

The data set of known NPs consists of a total of 208 166 unique compounds based on 50 366 unique Murcko scaffolds. It was compiled from (i) two encyclopedic NP databases and a general, smaller-sized NP database, (ii) four databases focused on NPs related to TCM, (iii) five databases focused on African plants and marine life, (iv) one database focused on Taiwanese plant species, (v) one database specializing in Brazilian species, and (vi) several others focused on specific biological activities and source organisms.

The data set of readily obtainable NPs consists of 25 524 unique compounds based on 5704 unique Murcko scaffolds. The majority of these NPs were retrieved from the overlap of the subset of readily obtainable compounds from ZINC 15^{27,28} (around 7.3 million compounds) with the set of all known NPs. In addition, we collected NPs from nine vendor libraries of readily obtainable NPs (see “physical NP libraries” in Table

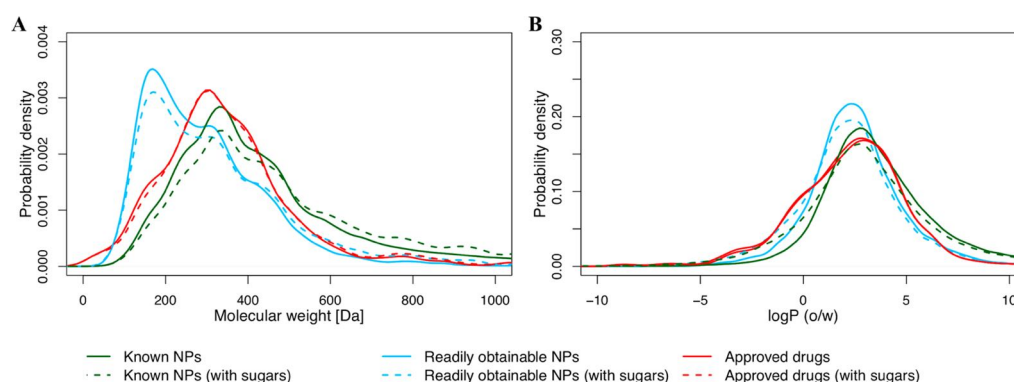


Figure 1. Distributions of (A) MW and (B) $\log P$ for all known NPs, readily obtainable NPs and approved drugs. Dashed lines indicate the distributions prior to the removal of sugars and sugar-like moieties (denoted as “sugars” in the figure legends); continuous lines indicate the distributions after the removal of these moieties.

1). These physical NP libraries are a subset of those discussed in our previous work⁷ that consist entirely of genuine NPs and for which we were able to obtain access to chemical information.

The chemical space of drugs is represented by the Approved Drugs subset of DrugBank.^{29,30} The data set consists of 1867 unique compounds based on 1053 Murcko scaffolds. In addition, on the basis of a comprehensive survey of natural products as sources of new drugs by Newman and Cragg,¹ we compiled a data set of 59 NPs and 320 NP derivatives (in this work denoted as the “Newman and Cragg data set”) that were approved as drugs between 1981 and 2014. Last but not least, we extracted a complete subset of known NPs cocrystallized with biomacromolecules from the Protein Data Bank (PDB)^{31,32} for analysis (see [Materials and Methods](#) for details).

Impact of the Removal of Sugars and Sugar-like Moieties on Physicochemical Properties. In the context of drug discovery, and drug design in particular, sugars and sugar-like moieties in NPs are generally not a focus of interest because they primarily affect pharmacokinetic properties and only rarely are essential for bioactivity. For this reason, we have devised and employed a set of rules (SMARTS patterns) to remove such moieties from the individual molecules (see [Materials and Methods](#) for details). The rule set identified sugars and sugar-like moieties in approximately 20% of all known and readily obtainable NPs (which is comparable to the rates reported in other studies^{3,16}) and in 5% of all approved drugs ([Table 1](#)).

In order to assess the impact of this processing step on this analysis, as a first step we compared the distributions of two key physicochemical properties, molecular weight (MW) and (calculated) $\log P$, prior to and after the removal of sugars and sugar-like moieties. As shown in [Figure 1](#), the procedure led to reductions in the median MWs of known and readily obtainable NPs by 43 and 22 Da, respectively, and increases in the median $\log P$ values by approximately 0.40 and 0.26 log units, respectively. The overall impact on approved drugs was very small ($\Delta\text{MW} = -4$ Da; $\Delta(\log P) = +0.04$ log units).

The further analysis is based on molecular structures after the removal of sugars and sugar-like moieties. The main text reports median or mean values as appropriate. Further data are reported in [Table 2](#) (the color gradient should not be interpreted as indicative of “better” or “worse” values).

Characterization of Known NPs, Readily Obtainable NPs, and Approved Drugs.

In order to determine the chemical space covered by known and readily obtainable NPs and compare it with that of approved drugs, principal component analysis (PCA) based on 17 relevant physicochemical properties was employed (see [Materials and Methods](#) for details). The score plots are reported in [Figure 2A](#); the loadings of the first and second principal components (PCs) are listed in [Table S1](#). PC1 and PC2 explain 39% and 14% of the total variance, respectively. There is no descriptor that dominates PC1 or PC2. Instead, features correlated with the size of a molecule, in particular MW, the number of heavy atoms, topological polar surface area, and number of hydrogen-bond acceptors are major contributors to PC1. The sum of formal charges, number of acidic atoms, and $\log P$ are major contributors to PC2.

The score plot shows that the chemical space covered by known NPs is substantially larger than that of readily obtainable NPs and drugs. The chemical space of drugs is well-represented by that of readily obtainable NPs. Approved drugs contain fewer phenols and phenol ethers (26%) than known and readily obtainable NPs (37% and 35%, respectively).

With respect to individual physicochemical properties, the overall distribution of MW (which is correlated with the size of molecules) among known NPs is similar to that of drugs ([Figure 1A](#)). The median MWs of known NPs and drugs are 381 and 326 Da, respectively. Unsurprisingly, there is a larger fraction of heavy compounds (MW greater than 500 Da) found among known NPs (27%) than among drugs (15%). Interestingly, however, readily obtainable NPs are substantially smaller than known NPs and drugs (median MW of 266 Da). More than 58% of readily obtainable NPs are fragment-sized (MW less than 300 Da), which is a much higher proportion than among known NPs (28%) and drugs (41%). Fragment-sized NPs can be of high value as starting points for optimization, which may explain their large share among readily obtainable NPs.

The median $\log P$ of known NPs (3.33) is almost 1 log unit higher than that of drugs (2.46), indicating that NPs are generally more hydrophobic ([Figure 1B](#)). In contrast, the median $\log P$ values of readily obtainable NPs and drugs are comparable (2.47 vs 2.46, respectively).

The molecular flexibility of molecules is in part represented by the number of rotatable bonds. The average numbers of

Table 2. Overview of the Physicochemical Properties of Known NPs, Readily Obtainable NPs, Drugs, and Individual Databases Investigated in this Work^a

	MW [Da]		Log P		No. of rot. bonds		Fraction of rot. bonds		No. of chiral centers		Fraction of C _{sp} ³ atoms		No. of rings		No. of aromatic rings		No. of N atoms		No. of O atoms		No. of H-bond accept.		No. of H-bond don.		No. of acidic atoms		No. of basic atoms		Lipinski's rule of five ^c	
	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	
Reference data sets																														
Known NPs	381	442	3.33	3.88	6.47	0.15	0.20	4.69	0.59	0.56	3.79	1.22	0.84	5.48	4.91	2.65	0.43	0.29	0.77											
Readily obtainable NPs	266	296	2.47	2.69	4.86	0.18	0.23	2.12	0.45	0.48	2.27	1.07	0.73	3.81	3.51	1.86	0.50	0.16	0.94											
Approved drugs	326	373	2.46	2.31	6.58	0.21	0.22	1.98	0.42	0.44	2.71	1.52	2.58	3.67	4.39	2.62	0.81	0.65	0.89											
Newman and Cragg data set ^b	434	674	2.21	1.25	16.34	0.23	0.27	6.39	0.60	0.59	3.62	1.31	5.29	8.65	9.06	6.95	1.50	1.03	0.66											
NPs of PDB	189	251	1.10	1.37	5.14	0.21	0.27	1.49	0.50	0.48	1.37	0.75	1.12	3.50	3.31	2.38	1.23	0.32	0.91											
Virtual NP databases: Encyclopedic and general NP databases																														
DNP	351	387	3.02	3.37	5.81	0.15	0.20	3.99	0.60	0.56	3.11	1.01	0.64	5.06	4.49	2.34	0.41	0.16	0.86											
UNPD	365	404	3.19	3.56	5.96	0.14	0.19	4.20	0.59	0.56	3.27	1.06	0.64	5.26	4.58	2.32	0.38	0.15	0.84											
NPs of PubChem Substance Database	320	327	2.56	2.53	4.68	0.17	0.18	1.54	0.33	0.36	3.01	1.83	1.37	4.07	3.84	1.90	0.53	0.14	0.97											
Virtual NP databases related to TCM																														
TCM Database@Taiwan	536	597	4.71	5.49	7.36	0.13	0.17	6.94	0.57	0.56	6.26	1.97	1.28	6.38	6.03	3.56	0.45	0.80	0.51											
TCMID	340	364	3.01	3.37	4.62	0.12	0.17	3.77	0.54	0.53	3.36	1.15	0.44	4.75	4.20	2.04	0.28	0.19	0.89											
HIM	282	280	2.18	2.27	3.05	0.12	0.14	1.51	0.30	0.34	2.61	1.50	0.49	4.15	4.09	2.38	0.54	0.17	0.96											
VIT	276	283	2.48	2.73	3.27	0.12	0.17	1.75	0.35	0.39	2.67	1.40	0.43	3.89	3.68	2.01	0.41	0.20	0.94											
Virtual NP databases focused on specific geographical regions and bioactivities																														
AfroDb	374	385	3.66	4.12	4.37	0.11	0.14	3.00	0.36	0.44	3.76	1.75	0.31	4.82	4.40	2.00	0.21	0.14	0.91											
AfroCancer	372	385	3.40	3.95	4.10	0.09	0.13	3.78	0.39	0.47	3.84	1.55	0.25	5.04	4.47	2.22	0.38	0.06	0.89											
AfroMalariaDB	362	363	3.48	3.69	3.54	0.11	0.12	2.59	0.35	0.40	3.71	1.74	0.32	4.79	4.34	1.95	0.16	0.15	0.95											
NANPDB	328	366	2.66	3.25	4.43	0.12	0.16	4.11	0.63	0.57	3.13	0.92	0.24	5.34	4.41	2.30	0.28	0.08	0.91											
SANCDDB	348	382	3.33	3.42	3.52	0.10	0.13	4.26	0.54	0.53	3.67	0.99	0.55	4.42	4.22	1.92	0.10	0.20	0.94											
TIPdb	336	361	2.98	3.37	4.30	0.12	0.16	3.16	0.43	0.48	3.37	1.41	0.34	5.03	4.44	2.08	0.29	0.12	0.89											
NuBBE	326	341	3.18	3.58	4.33	0.13	0.17	3.03	0.43	0.48	3.16	1.32	0.26	4.42	3.78	1.40	0.19	0.06	0.94											
UEFS Natural Products	316	334	3.12	3.64	3.24	0.10	0.13	3.29	0.50	0.50	3.33	1.23	0.27	4.07	3.67	1.77	0.29	0.11	0.94											
NPACT	382	398	3.58	3.94	5.41	0.12	0.16	4.11	0.47	0.50	3.59	1.32	0.17	5.26	4.63	2.19	0.24	0.06	0.83											
NPCARE	387	429	3.43	3.59	5.66	0.12	0.16	4.44	0.55	0.52	3.58	1.22	0.79	5.53	4.94	2.56	0.33	0.15	0.83											
StreptomeDB	396	459	2.15	2.10	6.78	0.15	0.20	4.57	0.50	0.49	2.90	1.18	2.13	6.60	6.59	4.36	0.98	0.34	0.70											
Physical NP libraries																														
Ambinter and Greenpharma NPs	333	365	2.88	3.09	4.11	0.12	0.15	3.46	0.43	0.47	3.35	1.36	0.62	4.89	4.43	2.22	0.37	0.16	0.90											
AnalytiCon Discovery MEGx	332	353	2.64	2.71	5.00	0.15	0.19	3.37	0.50	0.50	2.76	1.05	0.38	5.29	4.58	2.61	0.63	0.04	0.93											
Pi Chemicals NPs	337	363	2.75	2.87	4.04	0.11	0.14	3.62	0.45	0.47	3.44	1.33	0.56	5.05	4.60	2.44	0.41	0.21	0.90											
InterBioScreen NPs	309	324	2.42	2.63	3.26	0.10	0.13	3.87	0.54	0.54	3.29	1.04	0.84	3.92	3.84	1.87	0.41	0.34	0.95											
TargetMol Natural Compound Library	295	336	2.42	2.49	3.94	0.11	0.15	3.41	0.47	0.48	2.92	1.11	0.76	4.56	4.34	2.59	0.57	0.23	0.92											
p-ANAPL	330	342	2.81	3.37	2.99	0.10	0.11	1.83	0.21	0.29	3.35	1.96	0.22	4.97	4.57	2.26	0.21	0.08	0.92											
Developmental Therapeutic Program (DTP), NCI/NIH NP Set IV	349	391	2.31	2.29	3.59	0.09	0.12	3.89	0.46	0.48	3.70	1.35	1.32	5.54	5.43	2.41	0.42	0.32	0.84											
AK Scientific NPs	281	314	2.37	2.55	4.51	0.14	0.19	2.03	0.40	0.44	2.52	1.29	0.76	4.22	3.94	2.39	0.48	0.28	0.91											
Selleck Chemicals NPs	286	309	2.31	2.21	2.83	0.09	0.12	2.66	0.29	0.39	3.12	1.46	0.68	4.60	4.39	2.46	0.36	0.20	0.95											

^aValues are indicated by a color gradient, ranging from dark blue (minimum value) via white to dark green (maximum value). The color gradient is not indicative of "better" or "worse" values. ^bSubset of the Newman and Cragg data set containing all drugs approved between 1981 and 2014 that are NPs or NP derivatives. ^cFraction of compounds complying with Lipinski's rule of five.

rotatable bonds of known NPs and drugs are comparable (6.47 vs 6.58; Figure 3A), and so are their average fractions of rotatable bonds (0.20 vs 0.22; Figure 3B). As a result of the difference in molecular size, the average number of rotatable bonds of readily obtainable NPs is smaller (4.86), although their average fraction of rotatable bonds is similar (0.23).

Known NPs have on average more than twice as many chiral centers (4.69) as drugs (1.98) and readily obtainable NPs (2.12) (Figure 3C). The largest difference is observed for the fraction of achiral compounds, which is approximately 45% for

drugs and readily obtainable NPs but less than 25% for known NPs. When the removal of sugars and sugar-like moieties (which contain several stereogenic atoms) is taken into consideration, these findings are in line with previous reports.^{4,18,74,75}

A second characteristic number that reflects the three-dimensional complexity of molecular structures is the number of C_{sp}³ atoms, for which trends similar to those observed for chiral centers are recorded. The median fraction of C_{sp}³ atoms of known NPs (0.59) is higher than those of readily obtainable

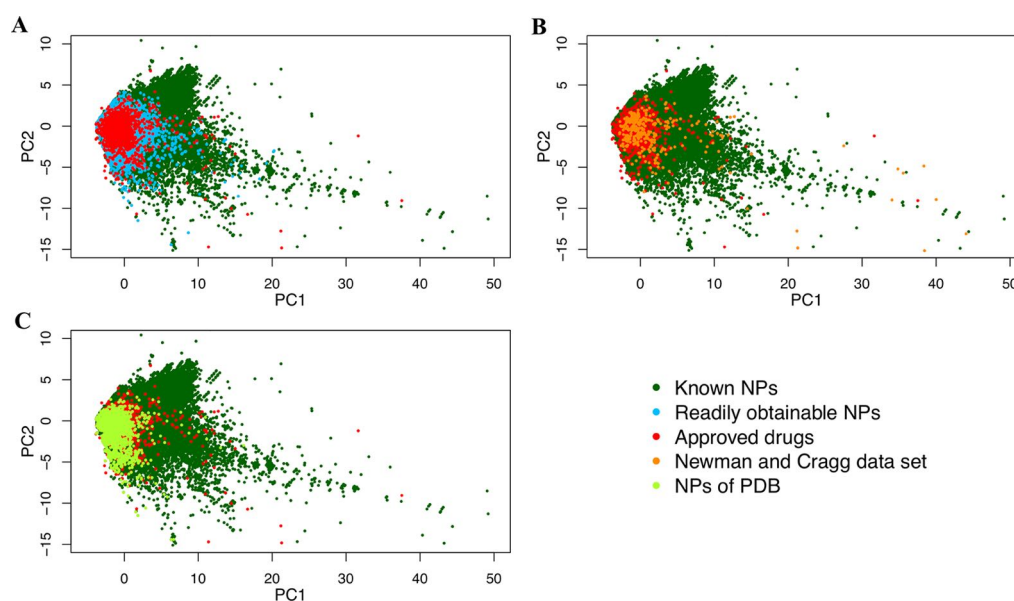


Figure 2. Scatter plots of the second PC against the first PC for the reference data sets based on 17 relevant physicochemical properties. All of the panels show the data points of the known NPs and approved drugs. In addition, the individual panels show (A) the readily obtainable NPs, (B) the Newman and Cragg data set, and (C) the NPs present in X-ray structures in the PDB. PC1 and PC2 explain 39% and 14% of the total variance, respectively.

NPs (0.45) and drugs (0.42), which is in agreement with previous reports.^{19,75} The density distribution plot (Figure 3D) shows that a large portion of drugs have a smaller fraction of C_{sp^3} atoms than NPs, with the highest densities recorded around a value of 0.4. For known NPs, the highest densities are observed around a value of 0.8.

Although known NPs on average consist of more rings (3.79) than drugs (2.71) (Figure 3E), known NPs contain fewer aromatic rings than drugs (1.22 vs 1.52; Figure 3F). Most noticeable is that readily obtainable NPs usually have no rings or just a single ring, whereas most drugs and known NPs have three or four rings. Many of these trends can be attributed to the size of molecules and are in agreement with results published elsewhere.^{4,16,19,74,76} More details on the propensities of various types of rings are provided in Figure S1.

Known NPs have on average approximately 70% fewer nitrogen atoms (Figure 3G) than drugs and 45% more oxygen atoms (Figure 3H) than drugs and readily obtainable NPs. The findings differ from previous reports (e.g., ref 4) because of the removal of sugars and sugar-like moieties. Related to the occurrence of oxygen and nitrogen atoms and size, a difference in the occurrence of hydrogen-bond acceptors and donors is observed among known NPs (4.91 and 2.65, respectively), readily obtainable NPs (3.51 and 1.86, respectively), and drugs (4.39 and 2.62, respectively). Most obvious in this regard is an accumulation of readily obtainable NPs with few hydrogen-bond acceptors and donors (Figure 3I,J).

Drugs have on average more acidic atoms (0.81) than known NPs (0.43) and readily obtainable NPs (0.50) (Figure 3K). Because of the higher propensity of nitrogen atoms in drugs, they also have on average more basic atoms (0.65) than known NPs (0.29) and readily obtainable NPs (0.16) (Figure 3L).

Approximately three-quarters of known NPs comply with Lipinski's rule of five.⁷⁷ Interestingly, the fraction of readily obtainable NPs satisfying Lipinski's rule of five (0.94) is higher

than the fraction of approved drugs that comply with the rule (0.89).

Characterization of Approved Drugs That Are NPs or NP Derivatives. The prepared Newman and Cragg data set consists of a total of 59 NPs and 320 NP derivatives that were approved as drugs between 1981 and 2014 (see [Materials and Methods](#) for details). Most of these compounds are located in regions of chemical space that are most densely populated with approved drugs (Figure 2B). Over one-third of these compounds are basic alkaloids, 62% are alkaloids according to the extended definition of "alkaloids" (see footnote *h* of Table 1 for details), 27% are phenols and phenol ethers (vs 26% for all approved drugs), and 12% are steroids (vs 5% for all approved drugs). The density distribution of the MW of these compounds (Figure 4A) shows an accumulation of heavy compounds (mostly peptides) among these drugs (median MW of 434 Da vs 326 Da for all approved drugs). Related to the greater size of these compounds, an exceptionally high number of rotatable bonds (mean 16.34 vs 6.58 for all approved drugs) and one of the highest numbers of chiral centers (mean 6.39 vs 1.98 for all approved drugs) are observed (Table 2). Also, the average numbers of nitrogen and oxygen atoms, hydrogen-bond donors and acceptors, and acidic and basic atoms are approximately twice as high as those of approved drugs. In contrast, the median log *P* (2.21) is comparable to that of approved drugs (2.46), but its distribution is wider (Figure 4B). Only two-thirds of all compounds in this data set comply with Lipinski's rule of five (vs 89% for all approved drugs).

Characterization of NPs Cocrystallized with Biomacromolecules. For 2060 unique NPs (based on 510 Murcko scaffolds), at least one X-ray crystal structure of good quality (according to the assessment of global quality measures described in [Materials and Methods](#)) in complex with a biomacromolecule is available. The PCA plot in Figure 2C shows a high density of cocrystallized ligands in areas of

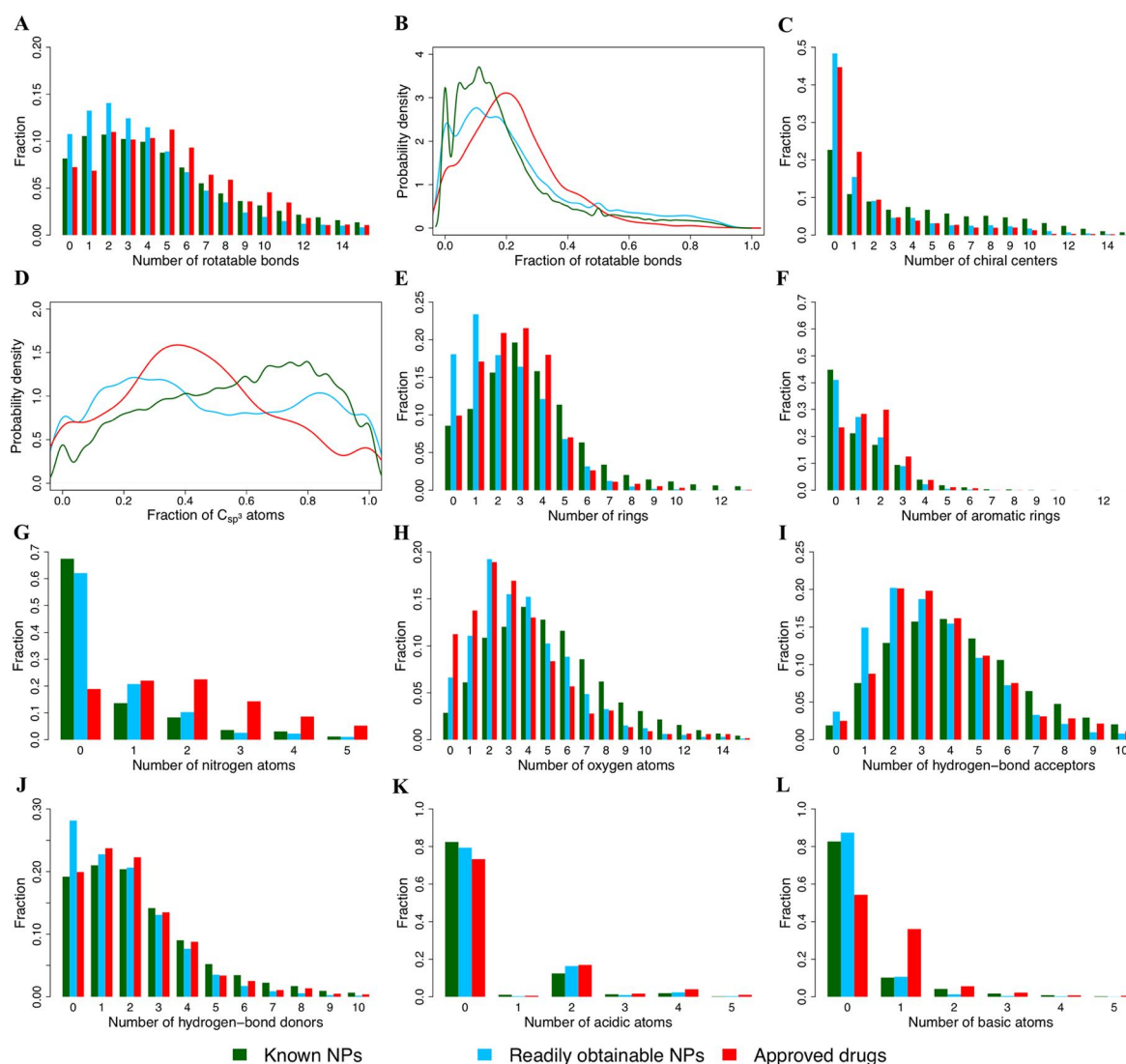


Figure 3. Distributions of key physicochemical properties among known NPs, readily obtainable NPs, and approved drugs: (A) number of rotatable bonds, (B) fraction of rotatable bonds, (C) number of chiral centers, (D) fraction of C_{sp^3} atoms, (E) number of rings, (F) number of aromatic rings, (G) number of nitrogen atoms, (H) number of oxygen atoms, (I) number of hydrogen-bond acceptors, (J) number of hydrogen-bond donors, (K) number of acidic atoms, and (L) number of basic atoms. Histograms of specific types of rings are provided in Figure S1.

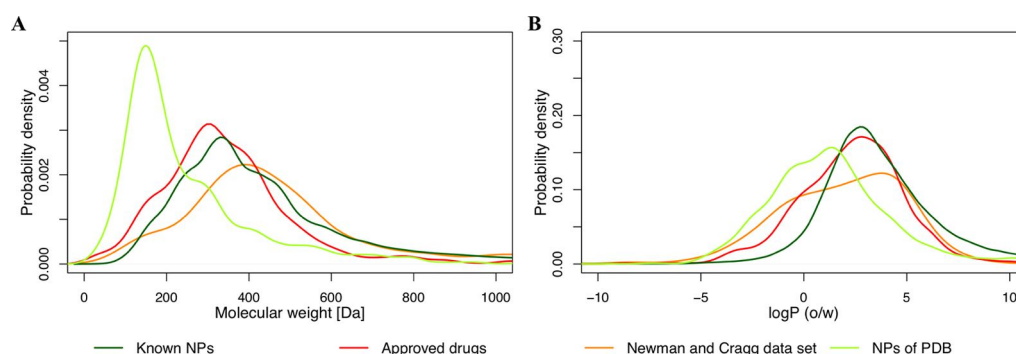


Figure 4. Distributions of (A) MW and (B) $\log P$ of known NPs, approved drugs, drugs that are NPs or derived therefrom (Newman and Cragg data set), and NPs for which at least one X-ray structure of a complex with a biomacromolecule has been deposited in the PDB.

chemical space densely populated with approved drugs but also a substantial number of compounds populating other areas.

Noteworthy is an accumulation of alkaloids (37% vs 26% for all known NPs) and a lower abundance of phenols and phenol

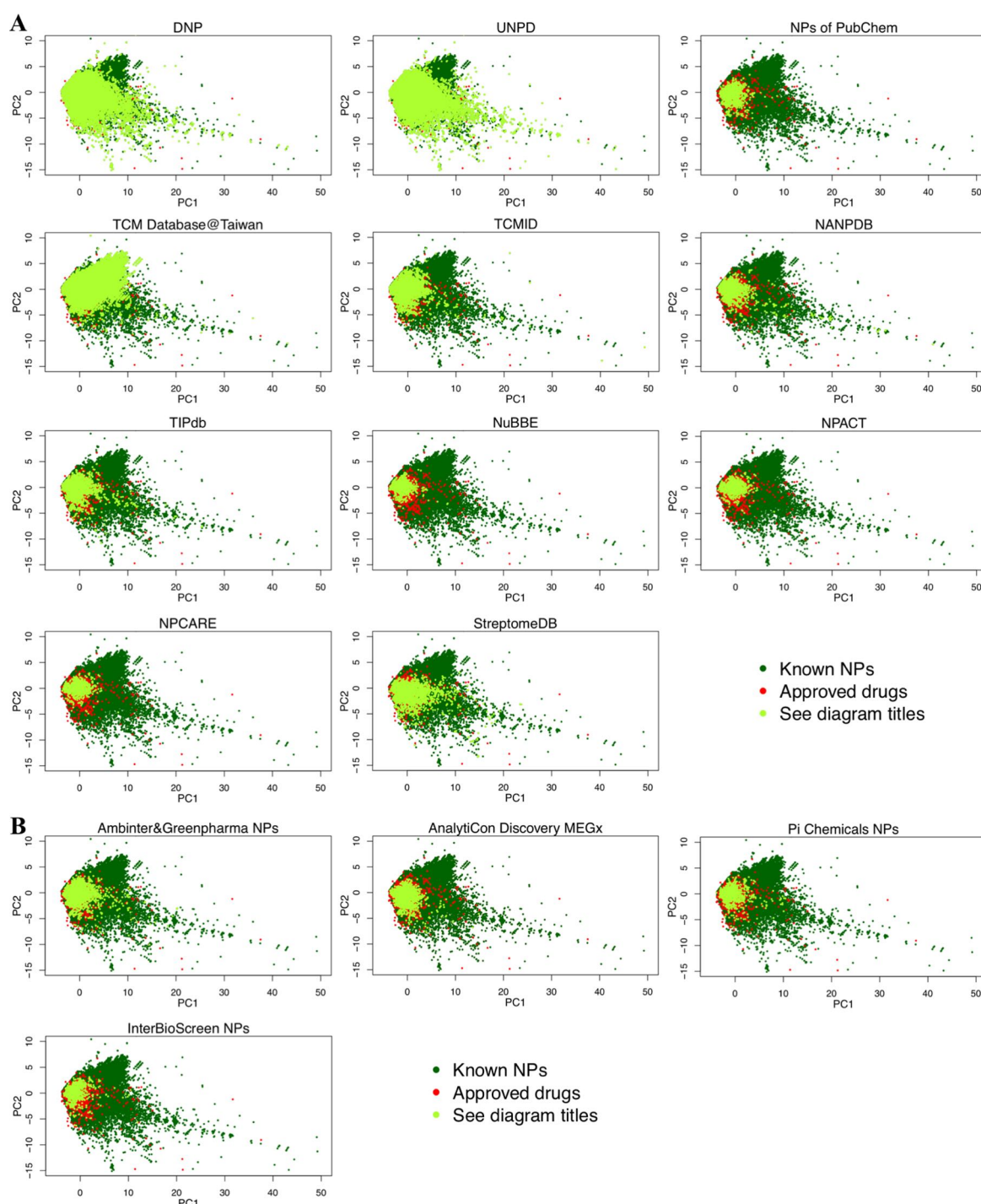


Figure 5. Scatter plots of the second PC against the first PC for (A) virtual NP databases and (B) physical NP libraries based on 17 relevant physicochemical properties. PC1 and PC2 explain 39% and 14% of the total variance, respectively.

ethers (22% vs 37% for all known NPs) (Table 1). Most cocrystallized NPs are of low MW (Figure 4A). The median is just 189 Da, compared with 326 Da for approved drugs and 381 Da for known NPs. However, there are several instances of high-MW NPs in this data set. In contrast to the Newman and Cragg data set, these NPs are not primarily peptides but rather macrolides, lipids, and oligosaccharides. The median log *P* is roughly 1 log unit lower than that of approved drugs and 2 log

units lower than that of known NPs (Figure 4B). In conjunction with the low average MW of cocrystallized NPs, in particular the numbers of rings (over 35% of these compounds are acyclic), oxygen atoms, and hydrogen-bond acceptors are lower than those of known NPs. Overall, with respect to the considered physicochemical properties, cocrystallized NPs are generally more similar to approved drugs than known NPs.

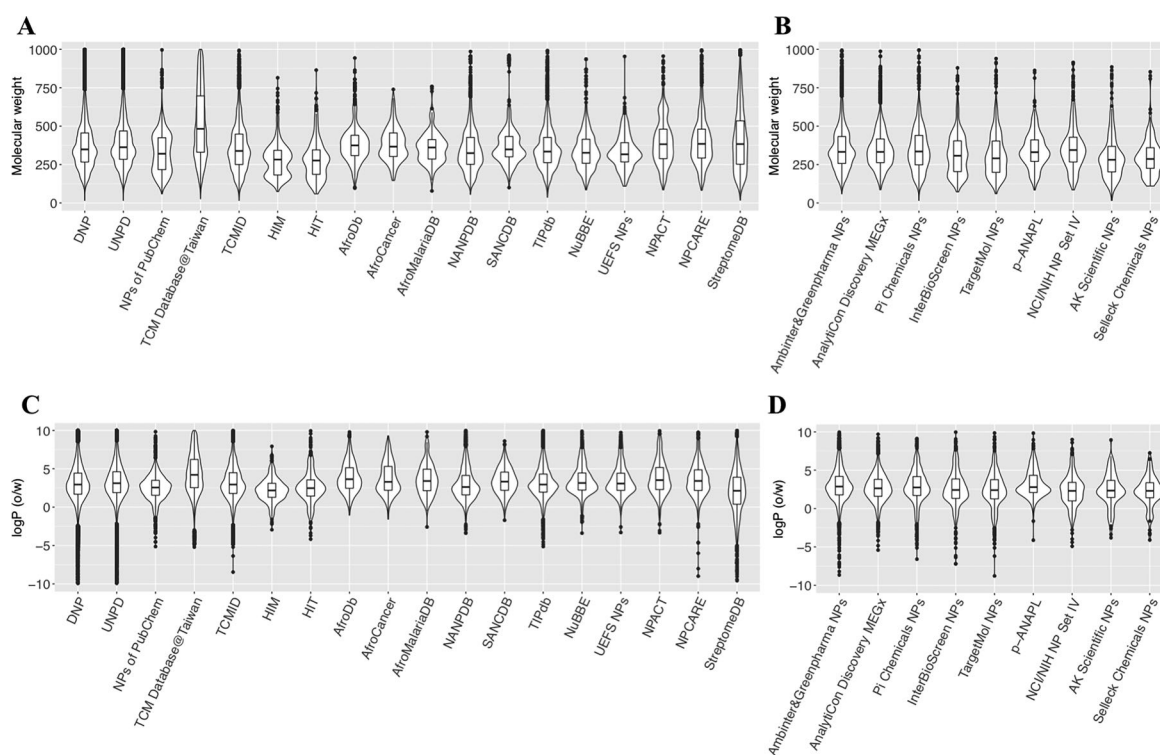


Figure 6. Violin and box plots illustrating the distributions of (A, B) MW and (C, D) $\log P$ for the (A, C) virtual NP databases and (B, D) physical NP libraries.

Characterization of Individual NP Databases. For all of the databases discussed in this section consisting of more than 1000 NPs, PCA plots illustrating the coverage of the chemical space are provided in Figure 5. Figures 6 and S2–S23 report the physicochemical property distributions of the individual databases. Statistical data are reported in Table 2.

Virtual NP Databases: Encyclopedic and General Databases. The DNP is the most established encyclopedic database of NPs and is used as the reference data set for this comparative analysis. It covers a total of more than 128 000 unique NPs based on over 28 000 different Murcko scaffolds (Table 1). Approximately 28% of these NPs are exclusively available via this virtual NP database (see the “exclusive NPs” columns in Table 1). The UNPD is slightly larger in size, consisting of over 140 000 unique NPs based on over 34 000 different Murcko scaffolds. Approximately one-third of these NPs are not covered by any other virtual NP database. Both DNP and UNPD clearly cover the largest regions of chemical space among all of the NP databases (Figure 5A), with all of the major NP classes well-represented (Table 1; it should be noted that NPs can be assigned to more than one scaffold class). The distributions of key physicochemical properties are similar for the two databases (Table 2 and Figures 6 and S2–S23): Approximately half of the compounds contained in the DNP and UNPD have a MW between 270 and 470 Da and a $\log P$ between 1.70 and 4.80. Approximately 85% of the NPs contained in either database comply with Lipinski’s rule of five.

The NPs subset of the PubChem Substance Database consists of 2760 unique NPs, about 40% of which are not included in any other virtual NP database. It is a diverse library of NPs (1031 Murcko scaffolds) that stands out because of its good coverage of the chemical space of approved drugs (Figure 5A). A strong presence of alkaloids (52% vs 20% for the DNP)

is notable, whereas the proportions of steroids (2% vs 7% for the DNP) and NPs containing sugars or sugar-like moieties (5% vs 17% for the DNP) are low. The database contains the highest proportion of NPs complying with Lipinski’s rule of five (97%). NPs from this library are on average smaller and less hydrophobic than those contained in the DNP (median MW and $\log P$ of 320 Da and 2.56, respectively, vs 351 Da and 3.02 for the DNP) but very comparable with approved drugs (median MW of 326 and $\log P$ of 2.46). They have on average the highest fraction of rotatable bonds (median of 0.17 vs 0.15 for the DNP) and second-highest number of aromatic rings (mean of 1.83 vs 1.01 for the DNP) among all of the virtual NP libraries.

Virtual NP Databases Related to TCM. Four databases containing NPs related to TCM are included in this analysis. The TCM Database@Taiwan is the most comprehensive database in this category by far, containing over 42 000 unique NPs based on ~16 000 Murcko scaffolds. More than half of the NPs in this database are not present in any other virtual NP library. The database is particularly rich in basic alkaloids (38% vs 11% for the DNP) and also in phenols and phenol ethers (51% vs 34% for the DNP). NPs from this database cover in part a unique region of chemical space (Figure 5A). The TCM Database@Taiwan stands out because of a substantially larger fraction of NPs with high MW (median of 536 Da vs 351 Da for the DNP), including in particular polyphenols and basic alkaloids. Also, the median $\log P$ for this data set (4.71) is much higher than those for the DNP (3.02) and any other database. This is associated with a large number of NPs with, e.g., many rotatable bonds (mean of 7.36 vs 5.81 for the DNP), rings (mean of 6.26 vs 3.11 for the DNP), and hydrogen-bond donors (mean of 3.56 vs 2.34 for the DNP) and acceptors (mean of 6.03 vs 4.49 for the DNP). Less than 15% of the NPs

included in this database are achiral, in contrast to 27% and 50% for the DNP and the NP subset of PubChem, respectively. Only 51% of the NPs contained in this database comply with Lipinski's rule of five, which is the lowest rate among all of the databases.

The TCMID focuses on the relationship between herbs, targets, and diseases and contains over 10 000 unique NPs representing more than 3600 Murcko scaffolds. However, most of the NPs contained in this database are also found in other virtual libraries. For example, the TCM Database@Taiwan contains 82% of the molecular structures in the TCMID. In contrast to the TCM Database@Taiwan, the physicochemical property distributions of the NPs contained in the TCMID are similar to those observed for the DNP. Also, the region of chemical space covered by this data set is more similar to that covered by approved drugs (Figure 5A).

HIM (636 unique NPs) and HIT (387 unique NPs) are two smaller-sized databases related to TCM. Around 27% of all NPs contained in HIM are exclusively available via this virtual NP library, whereas HIT is almost entirely covered by other resources. HIM provides metabolism information for herbal ingredients, whereas HIT connects herbal ingredients to their biological targets. The NPs of both databases have a lower median MW (282 and 276 Da, respectively) and log *P* (2.18 and 2.48, respectively) than those of the DNP (351 Da and 3.02, respectively). This is associated with, e.g., the lower numbers of rotatable bonds (means of 3.05 and 3.27), chiral centers (46% and 53% of NPs in HIM and HIT, respectively, are achiral), and oxygen and nitrogen atoms (fewer than 30% of the NPs in either database have at least one nitrogen atom). Approximately 95% of the NPs included in either database comply with Lipinski's rule of five.

Virtual NP Databases Focused on African Plants and Marine Life. AfroDb, AfroCancer, AfroMalariaDB, NANPDB, and SANCDB focus on NPs from plants and marine life indigenous to Africa. NANPDB contains 3297 unique NPs (1237 Murcko scaffolds) from Northern African species (mainly plants) and is the largest among these libraries. The other four databases contain several hundred NPs each. AfroDb is focused on NPs from African medicinal plants; SANCDB is a collection of NPs from South African plants and marine life; AfroCancer contains NPs from African medicinal plants with confirmed anticancer, cytotoxic, or antiproliferative activity, and AfroMalariaDB contains antimalarial NPs from African plant species. For AfroDb, AfroCancer, and AfroMalariaDB, an accumulation of phenolic NPs is observed (approximately 60% of the NPs from these databases contain at least one phenol or phenol ether). This is reflected in a high average number of aromatic rings (at least 1.55 vs 1.01 for the DNP) and low median fractions of C_{sp^3} atoms (at most 0.39 vs 0.60 for the DNP) and rotatable bonds (around 0.10 vs 0.15 for the DNP). With the exception of SANCDB, these databases contain a much lower percentage of alkaloids (around 10%) compared with the DNP (20%). This is associated with an equally lower average number of nitrogen atoms per NP (approximately 0.25–0.30) compared with the DNP (0.64). NANPDB stands out because it exhibits one of the highest rates of NPs containing sugars or sugar-like moieties (28% of NPs vs 17% for the DNP), whereas SANCDB exhibits one of the highest rates of steroids (14% of NPs vs 7% for the DNP) among all of the NP databases. The region of chemical space covered by NANPDB is similar to that covered by approved drugs (Figure 5A).

Other Virtual NP Databases with a Geographical Context. TIPdb is focused on anticancer, antituberculosis, and antiplatelet phytochemicals from plants indigenous to Taiwan. It contains 6673 unique NPs based on 2339 Murcko scaffolds that cover a region of chemical space similar to that of approved drugs (Figure 5A). Noteworthy are a below-average occurrence of alkaloids and low average number of nitrogen atoms (0.34 vs 0.64 for the DNP). No substantial differences in the physicochemical property distributions in comparison to the DNP are observed.

NuBBE focuses on NPs and derivatives from plants and microorganisms native to Brazil. It consists of 1613 unique NPs based on 636 Murcko scaffolds. The region of chemical space covered by this database is primarily a subspace of that of approved drugs (Figure 5A). NuBBE is characterized by a higher proportion of flavonoids (6% vs 2% for the DNP) and lower proportion of alkaloids (9% vs 20% for the DNP). The latter is reflected by one of the lowest numbers of nitrogen atoms (0.26 vs 0.64 for the DNP) and the lowest number of hydrogen-bond donors (1.40 vs 2.34 for the DNP) on average.

The UEFS Natural Products database contains 476 unique NPs based on 270 Murcko scaffolds. Notable is the very low rate of NPs containing sugars and sugar-like moieties (3% vs 17% for the DNP). The database contains a high proportion of flavonoids (9% vs 2% for the DNP) and few alkaloids (12% vs 20% for the DNP). Similar to NuBBE, the average numbers of nitrogen atoms (0.27) and hydrogen-bond donors (1.77) and acceptors (3.67) are low compared with those for the DNP (0.64, 2.34, and 4.49, respectively).

Virtual NP Databases Focused on Anticancer Activity. NPACT and NPCARE are smaller-sized databases focused on NPs related to anticancer activity. NPACT is characterized by a high proportion of phenols and phenol ethers (50% vs 34% for the DNP) and a low proportion of alkaloids (6% vs 20% for the DNP). This is accompanied by the lowest average number of nitrogen atoms (0.17 vs 0.64 for the DNP) among all databases (90% of the NPs contained in NPACT have no nitrogen atoms). NPCARE is more balanced than NPACT with respect to the NP classes. For both databases a strong accumulation of NPs in regions of chemical space densely populated with approved drugs is observed (Figure 5A).

Virtual NP Databases Focused on Specific Source Organisms. StreptomeDB consists of 3182 NPs based on 1322 Murcko scaffolds. Approximately one-third of the NPs listed in this database are not available from any other virtual NP library. The database is characterized by one of the highest proportions of NPs containing sugars or sugar-like moieties (25% vs 17% for the DNP). In the PCA plot, the majority of NPs of StreptomeDB are located in areas densely populated with approved drugs (Figure 5A). The database is rich in alkaloids (47% vs 20% for the DNP), although few of them contain a basic nitrogen atom (19% of all NPs contained in this data set). StreptomeDB contains only few steroids (1% vs 7% for the DNP). Notable are very broad distributions of MW (Figure 6A) and log *P* (Figure 6C). Also, higher abundances of nitrogen and oxygen atoms are observed, accompanied by an above-average number of moieties forming hydrogen bonds. Noteworthy is the low compliance with Lipinski's rule of five by NPs in this database (0.70 vs 0.86 for the DNP), which is undercut only by the TCM Database@Taiwan.

Physical NP Libraries. Score plots illustrating the regions of chemical space covered by four of the largest physical NP libraries are reported in Figure 5B. From the plots it can be

seen that the NPs of all of these libraries accumulate in areas densely populated with approved drugs.

Ambinter and Greenpharma NPs. Ambinter and Greenpharma offer a library of 6058 unique NPs based on 2331 Murcko scaffolds. The library contains more than 4000 NPs exclusively provided by this vendor, which is the largest number among all of the physical NP libraries investigated in this work. With respect to the prevalence of individual NP classes and physicochemical properties, this database is the one closest to the average of all NP databases. No substantial differences in the distributions of physicochemical properties compared to the DNP are observed.

AnalytiCon Discovery MEGx Database. The AnalytiCon Discovery MEGx database consists of 3346 unique NPs based on 1247 Murcko scaffolds that have been isolated from plants and microorganisms. The library stands out because of the largest percentage of NPs (73%) provided exclusively by this vendor. The database is also characterized by the highest proportion of NPs containing sugars or sugar-like moieties (35% vs 17% for the DNP). A below-average proportion of alkaloids is observed (14% vs 20% for the DNP; only 3% basic alkaloids), which is reflected by a low average number of nitrogen atoms (0.38 vs 0.64 for the DNP) and the lowest average number of basic atoms among all of the databases (0.04 vs 0.16 for DNP). Over 80% of NPs from this database do not contain a nitrogen atom.

Pi Chemicals NPs. Pi Chemicals offers 1244 unique NPs based on 561 Murcko scaffolds. The data set is characterized by the second highest proportion of NPs containing sugars or sugar-like moieties (32%) after the AnalytiCon Discovery MEGx database (35%). It contains an above-average percentage of steroids (10% vs 7% for the DNP). The distributions of most physicochemical properties are similar to those observed for the DNP.

InterBioScreen NPs. InterBioScreen offers a database of 1067 unique NPs based on 485 Murcko scaffolds. This database is characterized by the highest rate of steroids among all of the NP databases (15% vs 7% for the DNP) and one of the highest rates of alkaloids (35% vs 20% for the DNP). NPs from InterBioScreen are on average smaller (median MW of 309 Da) and more hydrophilic (median log *P* of 2.42) than NPs from the DNP (median MW of 351 Da and median log *P* of 3.02). They contain on average a much lower number of rotatable bonds (3.26) than the NPs from the DNP (5.81). More than half of the NPs from InterBioScreen (54%) have no more than three rotatable bonds. They have lower average numbers of oxygen atoms (3.92) and hydrogen-bond acceptors (3.84) than NPs from the DNP (5.06 and 4.49, respectively). Ninety-five percent of NPs from InterBioScreen comply with Lipinski's rule of five, which is the highest value among all of the physical databases (also reached by NPs from Selleck Chemicals).

TargetMol NPs. TargetMol provides an NP library of 674 unique NPs based on 319 Murcko scaffolds. A high percentage of NPs with sugar and sugar-like moieties is observed (27% vs 17% for the DNP). The NPs are smaller-sized (median MW of 295 Da vs 351 Da for the DNP) and more hydrophilic (median log *P* of 2.42 vs 3.02 for the DNP).

p-ANAPL. This physical NP library consists of 443 unique structures based on 191 Murcko scaffolds. Approximately two-thirds of these NPs are not available via any other physical NP library. Noteworthy is a prevalence of phenols and phenol ethers (72% vs 34% for the DNP) and flavonoids (11% vs 2%

for the DNP), whereas alkaloids are rare (6% vs 20% for the DNP). With respect to the physicochemical properties of NPs, the database stands out because it exhibits the second-lowest average number of rotatable bonds (2.99 vs 5.81 for the DNP) and the lowest median fraction of C_{sp³} atoms among all of the databases (0.21 vs 0.60 of DNP). Distinct compressed shapes of the violin plots of the fractions of rotatable bonds and C_{sp³} atoms are observed (Figure S3B). Furthermore, NPs from p-ANAPL have the lowest average number of chiral centers among all of the physical NP databases (1.83 vs 3.99 for the DNP). More than half of the NPs from p-ANAPL are achiral. The average number of nitrogen atoms is the lowest among all of the physical NP databases (0.22 vs 0.64 for the DNP). In contrast, NPs from p-ANAPL have a much higher average number of aromatic rings than those from most of the other databases (1.96 vs 1.01 for the DNP, the second highest overall). More than 80% of the molecules from p-ANAPL do not contain a nitrogen atom. Most of these observations are related to the abundance of phenols and phenol ethers in this database.

NCI/NIH NP Set IV. The NCI/NIH NP Set IV consists of 392 unique NPs representing 278 Murcko scaffolds. Approximately 61% of these NPs are not available from any other physical NP library. In terms of NP class distributions, the database is quite the opposite of p-ANAPL. It contains a high proportion of alkaloids (42% vs 20% for the DNP) and few flavonoids (1% vs 2% for the DNP). NPs from this data set have the highest median MW among all of the physical NP libraries (median MW of 349 Da) but are on average not particularly hydrophobic (median log *P* of 2.31 vs 3.02 for the DNP). Noteworthy is the lowest fraction of rotatable bonds among all of the databases (mean 0.12 vs 0.20 for the DNP). The large median MW and prevalence of alkaloids are reflected by high average numbers of rings (3.70 vs 3.11 for the DNP) and nitrogen atoms (1.32 vs 0.64 for the DNP). Only about 40% of NPs from the NCI/NIH NP Set IV are free of nitrogen atoms. More than 80% of the NPs from this data set comply with Lipinski's rule of five, which is the lowest value among all of the physical libraries.

AK Scientific NPs. AK Scientific offers 174 unique NPs based on 93 Murcko scaffolds. Noteworthy is the highest percentage of flavonoids among all of the databases (13% vs 2% for the DNP) and, related to this, a high proportion of phenols (40% vs 24% for the DNP). The average number of chiral centers is low (2.03 vs 3.99 for the DNP).

Selleck Chemicals NPs. Selleck Chemicals offers 155 unique NPs based on 92 Murcko scaffolds. Similar to the library from AK Scientific, this one features a high percentage of flavonoids (12%) and phenolic NPs (52%). Most obvious is the high rigidity of NPs from this library. The average number of rotatable bonds is just 2.83 (vs 5.81 for the DNP), which is the lowest value among all of the databases. Ninety-five percent of the NPs from Selleck Chemicals comply with Lipinski's rule of five.

CONCLUSIONS

In this work, we compiled comprehensive data sets of known and readily obtainable NPs to characterize their coverage of chemical space and compare it with that of approved drugs and various virtual and physical NP libraries. SugarBuster was developed as a new approach to remove sugars and sugar-like moieties from NPs because these moieties are rarely essential

to biological activity and, for this and other reasons, generally are not a focus of interest for drug discovery.

NPs cover a much wider region of chemical space than approved drugs, and a significant number of NPs are located in areas of chemical space that are also densely populated with drugs. This is established knowledge and part of the reason why NPs have been and remain one of the most productive sources of drug leads. At least 10% of the more than 250 000 known NPs for which chemical structures have been deposited in public databases are readily obtainable from commercial vendors and public research institutions. In previous work,⁷ we showed that at least 58 000 compounds that are structurally closely related to known NPs (and therefore considered putative NP analogues or derivatives) are readily obtainable.

This work shows that the readily obtainable NPs are highly diverse (representing more than 5700 different Murcko scaffolds) and cover the major NP classes. The vast majority of readily obtainable NPs share regions of chemical space with approved drugs. Nearly two-thirds of them are fragment-sized. All of these properties substantiate the high relevance of readily obtainable NPs to drug discovery. Interestingly, and relevant in particular to structure-based drug design, more than 2000 different NPs are represented by at least one X-ray crystal structure in complex with a biomacromolecule in the PDB. These NPs are generally smaller-sized and more hydrophilic than approved drugs.

A comprehensive analysis of drugs approved between 1981 and 2014 that are NPs or derived from NPs shows that the majority of these medicines are alkaloids (62%), followed by phenols and phenol ethers (27%) and steroids (12%). A significant number of these drugs (mostly peptides) have high MW.

We also characterized the chemical space covered by 18 virtual NP databases and nine physical NP libraries using the DNP as an encyclopedic reference. Several distinctive features of individual databases could be identified. For example, the NPs subset of the PubChem Substance Database stands out because of its high proportion of drug-like NPs and the TCM Database@Taiwan because of its coverage of a wide and in part unique region of chemical space containing many large and highly chiral NPs. In addition, p-ANAPL differs from all of the other databases by containing NPs with a very low fraction of C_{sp³} atoms and, associated with this, a very high number of aromatic rings.

Some NP databases are particularly rich in certain NP classes, such as the TCM Database@Taiwan in basic alkaloids, p-ANAPL in phenols and phenol ethers, and (the NP subset of) InterBioScreen in steroids. The physical NP libraries are characterized by smaller-sized and more hydrophilic, drug-like compounds that are located in regions of chemical space densely populated with approved drugs.

Overall, this work confirms the relevance of NPs as one of the most important sources of drug leads and provides a comprehensive and detailed view of known and readily obtainable NPs. We believe that these insights will be helpful in the selection of data sources for computer-guided drug discovery.

MATERIALS AND METHODS

Data Sets. The data set of known NPs was derived by combining all of the compounds of the virtual NP databases (see Table 1 for a complete list of databases). The data set of readily obtainable NPs was compiled from the physical NP

libraries (Table 1) and the NPs contained in ZINC. The latter data set was obtained by an InChI-based⁷⁸ (stereochemistry and fixed hydrogen layers disabled) overlap of all of the virtual NP databases with the “in stock” subset of ZINC 15.^{27,28}

Genuine NPs contained in the mixed physical compound library of Pi Chemicals were identified using the property tag “Natural or Semi-Natural”. In analogy, the property “index” was used to extract genuine NPs from the InterBioScreen Natural Compounds collection.

A comprehensive set of drugs that are NPs or NP derivatives and were approved between 1981 and 2014 were extracted from the Newman and Cragg data set (provided in the Supporting Information of ref 1). NPs and NP derivatives were identified by the “source” tags “N” and “ND”, respectively. Out of the 387 items assigned to these two categories, we were able to retrieve the chemical structures of 59 NPs and 320 NP derivatives from PubChem on the basis of the generic names provided by Newman and Cragg. Chemical structures could not be assigned without ambiguity to eight entries (most of which were mixtures). Those entries were removed from the data set.

NPs contained in the PDB were identified by overlapping the set of all known NPs with the complete set of small molecules extracted from structures stored in the PDB^{31,32} that match the following conditions: (i) has free ligand(s); (ii) experimental method is X-ray and has experimental data; (iii) R factor $R_{\text{work}} < 0.4$; (iv) R factor $R_{\text{free}} < 0.45$; (v) resolution $< 2.5 \text{ \AA}$.

Data Set Preparation. All of the molecules were neutralized, and hydrogens were added with the MOE⁷⁹ “Wash” node in KNIME.⁸⁰ During this process, the minor components of salts were removed. Sugars and sugar-like moieties were removed with a newly developed tool called SugarBuster. SugarBuster identifies and removes:

- five-membered aliphatic ring moieties with
 - exactly one heteroatom in the ring AND
 - all carbons forming the ring being a member of only one ring AND
 - at least two substituents with EITHER
 - two oxygen atoms attached directly to the ring OR
 - one oxygen atom and one nitrogen atom attached directly to the ring
- six-membered aliphatic ring moieties with
 - a maximum of one heteroatom in the ring AND
 - all carbons forming the ring being a member of only one ring AND
 - at least three substituents with EITHER
 - three oxygen atoms attached directly to the ring OR
 - two oxygen atoms and one nitrogen atom attached directly to the ring.

The aglycon component with the highest number of heavy atoms is returned after fragmentation. Following this procedure, the MOE “Wash” node was also used to produce copies of the molecular structures with charged strong acids and bases for the later calculation of descriptors that require charged molecules and for the identification of duplicate molecules on the basis of InChI notation (same procedure as described above). Any compound consisting of elements other than H, B, C, N, O, F, Si, P, S, Cl, Br, and I was removed. Molecular descriptors were calculated with RDKit⁸¹ and MOE.

Principal Component Analysis and Classification of Natural Products. Descriptors computed for PCA analysis were MW, log *P*, topological polar surface area (TPSA), number of hydrogen-bond acceptors (a_acc), number of hydrogen-bond donors (a_don), number of heavy atoms (a_heavy), fraction of rotatable bonds (b_rotR), number of nitrogen atoms (a_nN), number of oxygen atoms (a_nO), number of halogen atoms (Halogen), number of acidic atoms (a_acid), number of basic atoms (a_base), sum of formal charges (FCharge), number of aromatic atoms (a_aro), and number of chiral centers (chiral) calculated with MOE as well as the number of rings (NumRings) and fraction of C_{sp³} atoms (FractionC_{sp³}) calculated with RDKit.

The substructure matching methods implemented in RDKit were used to classify natural products as noted in the Table 1 footnotes.

■ ASSOCIATED CONTENT

📄 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00302.

Loadings of the first two components resulting from PCA analysis (Table S1); distributions of physicochemical properties for known NPs, readily obtainable NPs, virtual NP databases, physical NP libraries, the Newman and Cragg data set, and NPs in the PDB (Figures S1–S23) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kirchmair@zbh.uni-hamburg.de. Tel.: +49 (0) 40 42838 7303.

ORCID

Ya Chen: 0000-0001-5273-1815

Marina Garcia de Lomana: 0000-0002-9310-7290

Nils-Ole Friedrich: 0000-0002-8983-388X

Johannes Kirchmair: 0000-0003-2667-5877

Funding

Y.C. was supported by the China Scholarship Council (201606010345). M.G.d.L. was supported by a performance scholarship from Universität Hamburg and the Departmental Authority for Science, Research and Gender Equality of the Free and Hanseatic City of Hamburg.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Neann Mathai for discussion and proofreading the manuscript.

■ ABBREVIATIONS

AfroCancer, the African Anticancer Natural Products Library; AfroDb, NPs from African medicinal plants; AfroMalariaDB, the African Antimalarial Natural Products Library; DNP, Dictionary of Natural Products; HIM, the Herbal Ingredients in Vivo Metabolism Database; HIT, the Herbal Ingredients' Targets Database; MW, molecular weight; NANPDB, the Northern African Natural Products Database; ND, not determined; NP, natural product; NPACT, the Naturally Occurring Plant-Based Anticancer Compound-Activity-Target Database; NPCARE, Database of Natural Products for Cancer

Gene Regulation; NuBBE, Nuclei of Bioassays, Ecophysiology and Biosynthesis of Natural Products Database; p-ANAPL, the Pan-African Natural Products Library; PC, principal component; PCA, principal component analysis; PDB, Protein Data Bank; SANCDB, the South African Natural Compound Database; StreptomeDB, NPs produced by streptomycetes; TCM Database@Taiwan, the Traditional Chinese Medicine Database@Taiwan; TCMID, the Traditional Chinese Medicine Integrated Database; TIPdb, the Taiwan Indigenous Plant Database; UEFS Natural Products, the natural products database of the State University of Feira De Santana; UNPD, the Universal Natural Products Database.

■ REFERENCES

- (1) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79*, 629–661.
- (2) Singh, S. B.; Culbertson, C. J. Chemical Space and the Difference between Natural Products and Synthetics. In *Natural Product Chemistry for Drug Discovery*; Buss, A. D., Butler, M. S., Eds.; Chapter 2, pp 28–43.
- (3) Grabowski, K.; Baringhaus, K.-H.; Schneider, G. Scaffold Diversity of Natural Products: Inspiration for Combinatorial Library Design. *Nat. Prod. Rep.* **2008**, *25*, 892–904.
- (4) Ertl, P.; Schuffenhauer, A. Cheminformatics Analysis of Natural Products: Lessons from Nature Inspiring the Design of New Drugs. *Prog. Drug Res.* **2008**, *66*, 217–235.
- (5) Cragg, G. M.; Newman, D. J. Biodiversity: A Continuing Source of Novel Drug Leads. *Pure Appl. Chem.* **2005**, *77*, 7–24.
- (6) Rodrigues, T.; Reker, D.; Schneider, P.; Schneider, G. Counting on Natural Products for Drug Design. *Nat. Chem.* **2016**, *8*, 531–541.
- (7) Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* **2017**, *57*, 2099–2111.
- (8) Dictionary of Natural Products (DNP). <http://dnp.chemnetbase.com> (accessed April 7, 2017).
- (9) Chen, C. Y.-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening in Silico. *PLoS One* **2011**, *6*, e15939.
- (10) Ntie-Kang, F.; Zofou, D.; Babiaka, S. B.; Meudom, R.; Scharfe, M.; Lifongo, L. L.; Mbah, J. A.; Mbaze, L. M.; Sippl, W.; Efange, S. M. N. AfroDb: A Select Highly Potent and Diverse Natural Product Library from African Medicinal Plants. *PLoS One* **2013**, *8*, e78085.
- (11) Choi, H.; Cho, S. Y.; Pak, H. J.; Kim, Y.; Choi, J.-Y.; Lee, Y. J.; Gong, B. H.; Kang, Y. S.; Han, T.; Choi, G.; Cho, Y.; Lee, S.; Ryoo, D.; Park, H. NPCARE: Database of Natural Products and Fractional Extracts for Cancer Regulation. *J. Cheminf.* **2017**, *9*, 2.
- (12) Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J. The Re-Emergence of Natural Products for Drug Discovery in the Genomics Era. *Nat. Rev. Drug Discovery* **2015**, *14*, 111–129.
- (13) Reker, D.; Perna, A. M.; 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. Revealing the Macromolecular Targets of Complex Natural Products. *Nat. Chem.* **2014**, *6*, 1072–1078.
- (14) Rodrigues, T. Harnessing the Potential of Natural Products in Drug Discovery from a Cheminformatics Vantage Point. *Org. Biomol. Chem.* **2017**, *15* (44), 9275–9282.
- (15) Camp, D.; Garavelas, A.; Campitelli, M. Analysis of Physicochemical Properties for Drugs of Natural Origin. *J. Nat. Prod.* **2015**, *78*, 1370–1382.
- (16) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting Biologically Relevant Chemical Space: A Structural Classification of Natural Products (SCONP). *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17272–17277.
- (17) Stratton, C. F.; Newman, D. J.; Tan, D. S. Cheminformatic Comparison of Approved Drugs from Natural Product versus Synthetic Origins. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 4802–4807.

- (18) Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PLoS One* **2013**, *8*, e62839.
- (19) Chen, H.; Engkvist, O.; Blomberg, N.; Li, J. A Comparative Analysis of the Molecular Topologies for Drugs, Clinical Candidates, Natural Products, Human Metabolites and General Bioactive Compounds. *MedChemComm* **2012**, *3*, 312–321.
- (20) El-Elimat, T.; Zhang, X.; Jarjoura, D.; Moy, F. J.; Orjala, J.; Kinghorn, A. D.; Pearce, C. J.; Oberlies, N. H. Chemical Diversity of Metabolites from Fungi, Cyanobacteria, and Plants Relative to FDA-Approved Anticancer Agents. *ACS Med. Chem. Lett.* **2012**, *3*, 645–649.
- (21) Muigg, P.; Rosén, J.; Bohlin, L.; Backlund, A. In Silico Comparison of Marine, Terrestrial and Synthetic Compounds Using ChemGPS-NP for Navigating Chemical Space. *Phytochem. Rev.* **2013**, *12*, 449–457.
- (22) Shang, J.; Hu, B.; Wang, J.; Zhu, F.; Kang, Y.; Li, D.; Sun, H.; Kong, D.-X.; Hou, T. Cheminformatic Insight into the Differences between Terrestrial and Marine Originated Natural Products. *J. Chem. Inf. Model.* **2018**, *58* (6), 1182–1193.
- (23) Lucas, X.; Grüning, B. A.; Bleher, S.; Günther, S. The Purchasable Chemical Space: A Detailed Picture. *J. Chem. Inf. Model.* **2015**, *55*, 915–924.
- (24) Yongye, A. B.; Waddell, J.; Medina-Franco, J. L. Molecular Scaffold Analysis of Natural Products Databases in the Public Domain. *Chem. Biol. Drug Des.* **2012**, *80*, 717–724.
- (25) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.
- (26) Maybridge screening libraries. <https://www.maybridge.com> (accessed May 17, 2018).
- (27) ZINC15. <http://zinc15.docking.org> (accessed May 26, 2017).
- (28) Sterling, T.; Irwin, J. J. ZINC 15—Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
- (29) DrugBank, version 5.0.9. <https://www.drugbank.ca> (accessed Oct 25, 2017).
- (30) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; 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. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.
- (31) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (32) RCSB Protein Data Bank. <https://www.rcsb.org> (accessed Feb 12, 2018).
- (33) *Dictionary of Natural Products*, version 19.1; Chapman & Hall/CRC, 2010.
- (34) Universal Natural Products Database (UNPD). <http://pkuxj.pku.edu.cn/UNPD> (accessed Oct 17, 2016).
- (35) Hao, M.; Cheng, T.; Wang, Y.; Bryant, H. S. Web Search and Data Mining of Natural Products and Their Bioactivities in PubChem. *Sci. China: Chem.* **2013**, *56*, 1424–1435.
- (36) PubChem Substance. <http://ncbi.nlm.nih.gov/pcsubstance> (accessed April 7, 2017).
- (37) Traditional Chinese Medicine Database@Taiwan. <http://tcm.cmu.edu.tw> (accessed Oct 17, 2016).
- (38) Xue, R.; Fang, Z.; Zhang, M.; Yi, Z.; Wen, C.; Shi, T. TCMID: Traditional Chinese Medicine Integrative Database for Herb Molecular Mechanism Analysis. *Nucleic Acids Res.* **2013**, *41*, D1089–D1095.
- (39) TCMID. Traditional Chinese Medicine Integrated Database. www.megabionet.org/tcmid (accessed Oct 19, 2016).
- (40) Kang, H.; Tang, K.; Liu, Q.; Sun, Y.; Huang, Q.; Zhu, R.; Gao, J.; Zhang, D.; Huang, C.; Cao, Z. HIM-Herbal Ingredients in-Vivo Metabolism Database. *J. Cheminf.* **2013**, *5*, 28.
- (41) Herbal Ingredients In-Vivo Metabolism Database (HIM). <http://58.40.126.120:8080/him/> (accessed April 13, 2017).
- (42) Ye, H.; Ye, L.; Kang, H.; Zhang, D.; Tao, L.; Tang, K.; Liu, X.; Zhu, R.; Liu, Q.; Chen, Y. Z.; Li, Y.; Cao, Z. HIT: Linking Herbal Active Ingredients to Targets. *Nucleic Acids Res.* **2011**, *39*, D1055–D1059.
- (43) HIT. Herbal Ingredients' Targets Database. <http://lifecenter.sgst.cn/hit> (accessed April 13, 2017).
- (44) AfroDb. <http://african-compounds.org/about/afrodb> (accessed Oct 18, 2016).
- (45) Ntie-Kang, F.; Nwodo, J. N.; Ibezim, A.; Simoben, C. V.; Karaman, B.; Ngwa, V. F.; Sippl, W.; Adikwu, M. U.; Mbaze, L. M. Molecular Modeling of Potential Anticancer Agents from African Medicinal Plants. *J. Chem. Inf. Model.* **2014**, *54*, 2433–2450.
- (46) AfroCancer. <http://african-compounds.org/about/afrocancer> (accessed Feb 10, 2017).
- (47) Onguéné, P. A.; Ntie-Kang, F.; Mbah, J. A.; Lifongo, L. L.; Ndom, J. C.; Sippl, W.; Mbaze, L. M. 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.* **2014**, *4*, 6.
- (48) AfroMalariaDB. <http://african-compounds.org/about/afromalariadb> (accessed Feb 10, 2017).
- (49) Ntie-Kang, F.; Telukunta, K. K.; Döring, K.; Simoben, C. V.; Mouboko, A. F. A.; Malange, Y. I.; Njume, L. E.; Yong, J. N.; Sippl, W.; Günther, S. NANPDB: A Resource for Natural Products from Northern African Sources. *J. Nat. Prod.* **2017**, *80*, 2067–2076.
- (50) Northern African Natural Products Database (NANPDB) www.african-compounds.org/nanpdb (accessed April 5, 2017).
- (51) Hatherley, R.; Brown, D. K.; Musyoka, T. M.; Penkler, D. L.; Faya, N.; Lobb, K. A.; Tastan Bishop, Ö. SANCDB: A South African Natural Compound Database. *J. Cheminf.* **2015**, *7*, 29.
- (52) South African Natural Compounds Database (SANCDB). <http://sancdb.rubi.ru.ac.za> (accessed Feb 8, 2017).
- (53) Lin, Y.-C.; Wang, C.-C.; Chen, I.-S.; Jheng, J.-L.; Li, J.-H.; Tung, C.-W. TIPdb: A Database of Anticancer, Antiplatelet, and Antituberculosis Phytochemicals from Indigenous Plants in Taiwan. *Sci. World J.* **2013**, *2013*, 736386.
- (54) Tung, C.-W.; Lin, Y.-C.; Chang, H.-S.; Wang, C.-C.; Chen, I.-S.; Jheng, J.-L.; Li, J.-H. TIPdb-3D: The Three-Dimensional Structure Database of Phytochemicals from Taiwan Indigenous Plants. *Database* **2014**, *2014*, bau055.
- (55) Taiwan Indigenous Plant Database (TIPdb). <http://cwtung.kmu.edu.tw/tipdb> (accessed Oct 19, 2016).
- (56) Pilon, A. C.; Valli, M.; Dametto, A. C.; Pinto, M. E. F.; Freire, R. T.; Castro-Gamboa, I.; Andricopulo, A. D.; Bolzani, V. S. NuBBE: An Updated Database To Uncover Chemical and Biological Information from Brazilian Biodiversity. *Sci. Rep.* **2017**, *7*, 7215.
- (57) Núcleo de Bioensaios, Biossíntese e Ecofisiologia de Produtos Naturais (NuBBE). <http://nubbe.iq.unesp.br/portal/nubbedb.html> (accessed April 19, 2017).
- (58) Mangal, M.; Sagar, P.; Singh, H.; Raghava, G. P. S.; Agarwal, S. M. NPACT: Naturally Occurring Plant-Based Anti-Cancer Compound-Activity-Target Database. *Nucleic Acids Res.* **2013**, *41*, D1124–D1129.
- (59) Naturally Occurring Plant Based Anticancerous Compound-Activity-Target Data Base (NPACT). <http://crdd.osdd.net/raghava/npact> (accessed April 13, 2017).
- (60) Database of Natural Products for Cancer Gene Regulation (NPCARE). <http://silver.sejong.ac.kr/npcare> (accessed Feb 20, 2017).
- (61) Klementz, D.; Döring, K.; Lucas, X.; Telukunta, K. K.; Erxleben, A.; Deubel, D.; Erber, A.; Santillana, I.; Thomas, O. S.; Bechthold, A.; Günther, S. StreptomeDB 2.0—an Extended Resource of Natural Products Produced by Streptomycetes. *Nucleic Acids Res.* **2016**, *44*, D509–D514.
- (62) StreptomeDB. <http://132.230.56.4/streptomedb2> (accessed April 13, 2017).
- (63) Ambinter. www.ambinter.com (accessed June 2, 2017).
- (64) GreenPharma. www.greenpharma.com (accessed June 2, 2017).

- (65) AnalytiCon Discovery. www.ac-discovery.com (accessed Nov 14, 2017).
- (66) PI Chemicals. www.pipharm.com (accessed May 5, 2017).
- (67) InterBioScreen. www.ibscreen.com (accessed Nov 14, 2017).
- (68) TargetMol. www.targetmol.com (accessed May 17, 2017).
- (69) Ntie-Kang, F.; Amoa Onguéné, P.; Fotso, G. W.; Andrae-Marobela, K.; Bezabih, M.; Ndom, J. C.; Ngadjui, B. T.; Ogundaini, A. O.; Abegaz, B. M.; Meva'a, L. M. Virtualizing the p-ANAPL Library: A Step towards Drug Discovery from African Medicinal Plants. *PLoS One* **2014**, *9*, e90655.
- (70) NCI/NCH NP Set IV is available at the following website: NCI/NIH Developmental Therapeutics Program (DTP). Available Plates. http://dtp.cancer.gov/organization/dscb/obtaining/available_plates.htm (accessed Oct 20, 2016).
- (71) AK Scientific. www.aksci.com (accessed April 19, 2017).
- (72) Selleck Chemicals. www.selleckchem.com (accessed Nov 14, 2017).
- (73) Hänsel, R.; Sticher, O. *Pharmakognosie - Phytopharmazie*; Springer: Berlin, 2009.
- (74) Wetzel, S.; Schuffenhauer, A.; Roggo, S.; Ertl, P.; Waldmann, H. Cheminformatic Analysis of Natural Products and Their Chemical Space. *Chimia* **2007**, *61*, 355–360.
- (75) López-Vallejo, F.; Giulianotti, M. A.; Houghten, R. A.; Medina-Franco, J. L. Expanding the Medicinally Relevant Chemical Space with Compound Libraries. *Drug Discov. Drug Discovery Today* **2012**, *17*, 718–726.
- (76) Bon, R. S.; Waldmann, H. Bioactivity-Guided Navigation of Chemical Space. *Acc. Chem. Res.* **2010**, *43*, 1103–1114.
- (77) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.
- (78) *InChI*, version 1.05; IUPAC: Research Triangle Park, NC, 2017.
- (79) *Molecular Operating Environment (MOE)*, version 2016.08; Chemical Computing Group: Montreal, QC, 2016.
- (80) Berthold, M. R.; Cebon, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinel, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization*; Baier, D., Critchley, F., Decker, R., Diday, E., Greenacre, M., Lauro, C. N., Meulman, J., Monari, P., Nishisato, S., Ohsumi, N., Opitz, O., Ritter, G., Schader, M., Eds.; Springer: Berlin, 2007; pp 319–326.
- (81) RDKit: Open-Source Cheminformatics, version 2017.09.1, 2017. <http://www.rdkit.org>.

3.2. Machine Learning Method for Assessing Natural Product-Likeness

During our analysis of data resources for NP research we noticed that in some databases NPs are mixed with (semi-) synthetic compounds. On the other hand, we found some libraries of genuine NPs to be contaminated with NP-derivatives, NP-analogs and, in some cases, even synthetic reactants. There are also many libraries that contain valuable NPs, but do not mention the inclusion of such. We therefore concluded that a method for the automated identification of NPs and NP-like compounds would be desirable and of significance to NP-based drug discovery.

As part of this work we developed NP-Scout (D5), a set of random forest classifiers able to discriminate NPs (and NP-like compounds) from synthetic compounds with high accuracy. This method is built on updated collections of freely available NPs and an equivalent number of synthetic compounds. NP-Scout is accessible via the NERDD (New E-Resource for Drug Discovery) web server (<https://nerdd.zbh.uni-hamburg.de/>) which provides in silico tools for early drug discovery developed by our group (A1).

The details of this method are described in the following publication. In July 2020, the publication was selected by the editors as a *hot paper* (Editor's choice) of the journal *Biomolecules*.

[D5] Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **2019**, *9* (2), 43.

Available at <https://doi.org/10.3390/biom9020043>.

Y. Chen and J. Kirchmair conceptualized the work. Y. Chen collected, curated and prepared the data. Y. Chen developed the machine learning models and the web server, a task for which she received guidance from C. Stork. Y. Chen implemented the similarity maps, a task for which she received guidance from S. Hirte. Y. Chen wrote the largest part of the manuscript. J. Kirchmair supervised this work.

Reprinted from:

Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **2019**, *9* (2), 43.

This is an open access article licensed under a Creative Commons Attribution 4.0 International License.

The supplementary information of this article can be found in [Appendix B](#).

Article

NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules

Ya Chen ¹, Conrad Stork ¹, Steffen Hirte ¹ and Johannes Kirchmair ^{1,2,3,*}

¹ Center for Bioinformatics (ZBH), Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany; chen@zbh.uni-hamburg.de (Y.C.); stork@zbh.uni-hamburg.de (C.S.); steffen.hirte@studium.uni-hamburg.de (S.H.)

² Department of Chemistry, University of Bergen, 5007 Bergen, Norway

³ Computational Biology Unit (CBU), Department of Informatics, University of Bergen, 5008 Bergen, Norway

* Correspondence: johannes.kirchmair@uib.no or kirchmair@zbh.uni-hamburg.de; Tel.: +47-5558-3464

Received: 4 December 2018; Accepted: 21 January 2019; Published: 24 January 2019



Abstract: Natural products (NPs) remain the most prolific resource for the development of small-molecule drugs. Here we report a new machine learning approach that allows the identification of natural products with high accuracy. The method also generates similarity maps, which highlight atoms that contribute significantly to the classification of small molecules as a natural product or synthetic molecule. The method can hence be utilized to (i) identify natural products in large molecular libraries, (ii) quantify the natural product-likeness of small molecules, and (iii) visualize atoms in small molecules that are characteristic of natural products or synthetic molecules. The models are based on random forest classifiers trained on data sets consisting of more than 265,000 to 322,000 natural products and synthetic molecules. Two-dimensional molecular descriptors, MACCS keys and Morgan2 fingerprints were explored. On an independent test set the models reached areas under the receiver operating characteristic curve (AUC) of 0.997 and Matthews correlation coefficients (MCCs) of 0.954 and higher. The method was further tested on data from the Dictionary of Natural Products, ChEMBL and other resources. The best-performing models are accessible as a free web service at <http://npscout.zbh.uni-hamburg.de/npscout>.

Keywords: natural products; natural product-likeness; machine learning; random forest; classification; similarity maps; visualization; molecular fingerprints; web service

1. Introduction

Natural products (NPs) continue to be the most prolific resource for drug leads [1–4]. A recent analysis found that over 60% of all small-molecule drugs approved between 1981 and 2014 are genuine NPs, NP analogs or their derivatives, or compounds containing an NP pharmacophore [5]. NPs are characterized by enormous structural and physicochemical diversity [6–8]. Some of the regions in chemical space covered by NPs are not, or only rarely, populated by synthetic molecules (SMs) [7,9]. The structural complexity of many NPs exceeds that of compounds found in conventional synthetic libraries for screening, in particular with respect to stereochemical aspects, molecular shape, and ring systems [10–18].

The primary bottleneck of NP research is the scarcity of materials for testing. In a recent study, we showed that the molecular structures of more than 250,000 NPs have been deposited in public databases, and that only approximately 10% of these are readily obtainable from commercial providers and other sources [19].

Given the fact that NPs exhibit a wide range of biological activities that are of immediate relevance to human health, new avenues that would make NP research more effective are being explored, in particular, research involving computational approaches [2]. For example, computational methods have been employed successfully for the identification of bioactive NPs [20–22] and their bio-macromolecular targets [23–26]. They have also been successfully utilized for the design of simple synthetic, bioactive mimetics of NPs [27–29]. In this context, computational methods for quantifying the NP-likeness of compounds can be valuable tools to guide the de novo generation of NP mimetics and optimize the NP-likeness of lead compounds. Such methods may also be useful for identifying genuine NPs in commercial compound libraries, which often also contain SMs [19]. This can be valuable in the context of library design and for the prioritization of compounds for experimental testing.

The best-known in-silico approach for identifying NPs is the NP-likeness score developed by Ertl et al. [30]. The NP-likeness score is a Bayesian measure that quantifies a compound's similarity with the structural space of NPs based on structural fragments. As such, the model can identify sub-structures characteristic to NPs. The method has been re-implemented, with some modifications, in various platforms (e.g., [31–33]). Among them is the Natural-Product-Likeness Scoring System [31], which allows the calculation of the NP-likeness score (with some modifications). The Natural-Product-Likeness Scoring System also allows the use of customized data sets for training. An alternative approach for quantifying NP-Likeness, following a similar modeling strategy, but based on extended connectivity fingerprints (ECFPs), was reported by Yu [34]. Also a rule-based approach has been reported [35].

In this work, we present the development and validation of new machine learning models for the discrimination of NPs and SMs. To the best of our knowledge, these models are trained on the largest collection of known NPs that have been employed for the development of such classifiers. Among further developments, we present the utilization of similarity maps [36] for the visualization of atoms of a molecule, which are characteristic for NPs or SMs, according to the models.

2. Materials and Methods

2.1. Data Preparation

NPs were compiled from several physical and virtual NP databases (see Results for details). The chemical structures were parsed directly from SMILES notation, where available. Alternatively, chemical structures stored in chemical table files (e.g., SDF) were parsed with RDKit [37] and converted into SMILES. Minor components of salts were removed by the method described in ref. [38]. Any compounds with a molecular weight below 150 Da or above 1500 Da, and any compounds consisting of elements other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, or I were filtered. The “canonicalize” method, which was implemented in the “tautomer” class of MolVS [39], was used for neutralizing the molecular structures and merging tautomers. After the removal of duplicate SMILES (ignoring stereochemistry), the processed NP reference data set consisted of a total of 201,761 NPs.

SMs were compiled from the “in-stock” subset of ZINC [40,41]. In a first step, 500,000 compounds of ZINC were picked by random selection from the complete “in-stock” subset and pre-processed following the identical protocol used for the NP databases. After generating unique, canonicalized SMILES, any molecules present in the NP reference data set were removed from the SM data set (as determined by the comparison of canonicalized SMILES). Then, random sampling was used to compile a reference data set of SMs of identical size as the NP reference data set (i.e., 201,761 compounds).

The Dictionary of Natural Products (DNP) [42] and the ChEMBL database [43,44] were pre-processed following the identical protocol outlined for the NP and SM data sets. The ChEMBL sub-set of molecules, published in the Journal of Natural Products, was retrieved directly from ChEMBL [43,45]. The natural products subset of ZINC was downloaded from the ZINC website [46].

2.2. Principal Component Analysis

Fifteen two-dimensional molecular descriptors calculated with the Molecular Operating Environment (MOE) [47] were used for principle component analysis (PCA): MW (Weight), $\log P$ ($\log P$ (o/w)), topological polar surface area (TPSA), number of hydrogen bond acceptors (a_acc), number of hydrogen bond donors (a_don), number of heavy atoms (a_heavy), fraction of rotatable bonds (b_rotR), number of nitrogen atoms (a_nN), number of oxygen atoms (a_nO), number of acidic atoms (a_acid), number of basic atoms (a_base), sum of formal charges (FCharge), number of aromatic atoms (a_aro) and number of chiral centers (chiral), and number of rings (rings).

2.3. Model Building

Prior to model building, the preprocessed NP and SM reference data sets were merged, resulting in a total of 403,522 data records. The merged data set was then randomly split into a training set of 322,817 and a test set of 80,705 compounds (ratio of 4:1). In fingerprint space, structurally distinct molecules may have identical fingerprints. For this reason, de-duplication, based on fingerprints, was separately performed for all NPs and all SMs in the training data. Any fingerprints present in both the NP and SM subsets were removed, in order to avoid conflicting class labels. This procedure resulted in a training set of 156,119 NPs and 161,378 SMs represented by Morgan2 fingerprints, and in a training set of 108,393 NPs and 157,162 SMs represented by MACCS keys.

Morgan2 fingerprints (1024 bits) [48,49] and MACCS keys (166 bits) were calculated with RDKit, and 206 two-dimensional physicochemical property descriptors were calculated with MOE. Random forest classifiers (RFCs) were generated with scikit-learn [50,51] using default settings, except for “n_estimators”, which was set to “100”, and “class_weight”, which was set to “balanced”.

The NP-likeness calculator [30,31,52] was trained on atom signatures derived from the identical NP and SM data sets, used for training the RFCs. Subsequently, the NP-likeness score was calculated for each molecule in the test set, according to the atom signatures. All calculations used a signature height of 3, resulting in scores ranging from -3 to 3 . Molecules with a score greater than 0.0 were labeled as NPs, and molecules with a score lower, or equal to 0.0 were labeled as SMs. NP class probabilities (and AUCs) were derived by normalizing these scores to a range from 0.0 to 1.0 .

2.4. Similarity Maps

Similarity maps were computed with the RDKit [37] Chem.Draw.SimilarityMaps module based on RFCs derived from Morgan2 fingerprints (1024 bits).

3. Results

3.1. Compilation of Data Sets for Model Development

An NP reference data set of 201,761 unique NPs was compiled from 18 virtual NP libraries and nine physical NP databases. The reference data set is identical to that compiled as part of our previous work [8], with two amendments: First, the compounds of the DNP [42] were not included in the data set, as they serve as an external test set in this work, and second, the recently published Natural Products Atlas database [53] was added as a new data source. An overview of the NP data sources utilized in this work is provided in Table 1. The table also reports the number of molecules that are contained in the individual databases prior to, and after, data preprocessing. This is a procedure that includes the removal of salt components and stereochemical information, the filtering of molecules composed of uncommon elements, and with a molecular weight (MW) below 150 Da or above 1500 Da, and the removal of duplicate molecules (see Methods for details). An equal amount (i.e., 201,761) of synthetic organic molecules (SMs) was collected from the “in-stock” subset of ZINC [41] by random selection.

Table 1. Size of the individual data sets prior to and after data preprocessing.

Name ¹	Number of Molecules in SMILES Notation Successfully Parsed with RDKit	Number of Unique Molecules After Data Preprocessing	Scientific Literature and/or Online Presence
UNPD	229,140	161,228	[54,55]
TCM Database@Taiwan	56,325	45,422	[56,57]
NP Atlas	20,018	18,358	[53]
TCMID	13,188	10,918	[58,59]
TIPdb	8838	7620	[60–62]
Ambinter and Greenpharma NPs	7905	6680	[63,64]
AnalytiCon Discovery MEGx	4315	4063	[65]
NANPDB	6841	3734	[66,67]
StreptomeDB	3990	3353	[68,69]
NPs of PubChem Substance Database	3533	2638	[70,71]
NuBBE	1856	1637	[72,73]
Pi Chemicals NPs	1783	1511	[74]
NPCARE	1613	1479	[75,76]
NPACT	1516	1376	[77,78]
InterBioScreen NPs	1359	1116	[79]
AfroDb	954	865	[80,81]
TargetMol Natural Compound Library	850	745	[82]
HIM	1284	641	[83,84]
SANCDDB	623	588	[85,86]
UEFS Natural Products	493	469	via ZINC [40,87]
p-ANAPL	538	456	[88]
NCI/NIH DTP NP set IV	419	394	[89]
HIT	707	362	[90,91]
AfroCancer	388	352	[92,93]
AfroMalariaDB	265	250	[94,95]
AK Scientific NPs	242	177	[96]
Selleck Chemicals NPs	173	163	[97]
NP data set TOTAL	-	201761	

¹ UNPD: the Universal Natural Products Database; TCM Database@Taiwan: the Traditional Chinese Medicine Database@Taiwan; NP Atlas: the Natural Products Atlas; TCMID: the Traditional Chinese Medicine Integrated Database; TIPdb: the Taiwan Indigenous Plant Database; NANPDB: the Northern African Natural Products Database; StreptomeDB: Streptome Database; NuBBE: Nuclei of Bioassays, Ecophysiology and Biosynthesis of Natural Products Database; NPCARE: Database of Natural Products for Cancer Gene Regulation; NPACT: the Naturally Occurring Plant-based Anti-Cancer Compound-Activity-Target Database; AfroDb: NPs from African medicinal plants; HIM: the Herbal Ingredients in-vivo Metabolism Database; UEFS Natural Products: the natural products database of the State University of Feira De Santana; p-ANAPL: the Pan-African Natural Products Library; NCI/NIH DTP NP set IV: the NP (plated) set IV of the Developmental Therapeutic Program of the National Cancer Institute/National Institutes of Health; HIT, the Herbal Ingredients' Targets Database; AfroCancer, the African Anticancer Natural Products Library; AfroMalariaDB, the African Antimalarial Natural Products Library.

3.2. Analysis of the Physicochemical Properties of Natural Products and Synthetic Molecules

Prior to model development, we compared the chemical space covered by the 201,761 unique NPs, and the equal number of unique SMs, using principal component analysis (PCA), based on 15 relevant physicochemical properties (see Methods for details). The score plot in Figure 1 shows that the chemical space of SMs is essentially a sub-space of NPs.

NPs have on average a higher MW than SMs (506 Da vs 384 Da) and a larger proportion of heavy compounds (38% vs. 10% of all molecules have a MW greater than 500 Da; Figure 2a). SMs have a narrower distribution of calculated log *P* values as compared to NPs (Figure 2b) but their averages are comparable (3.31 versus 3.25). SMs and NPs show clear differences in the entropy of element distributions in molecules, with NPs having, on average, a lower entropy than SMs (1.39 versus 1.63; Figure 2c). NPs tend to have more chiral centers (mean 6.66 vs. 0.75; Figure 2d), substantially fewer nitrogen atoms than SMs (mean 0.76 vs. 2.94; Figure 2e), and more oxygen atoms (mean 7.39 vs. 2.88; Figure 2f) [7,10,12–15,17].

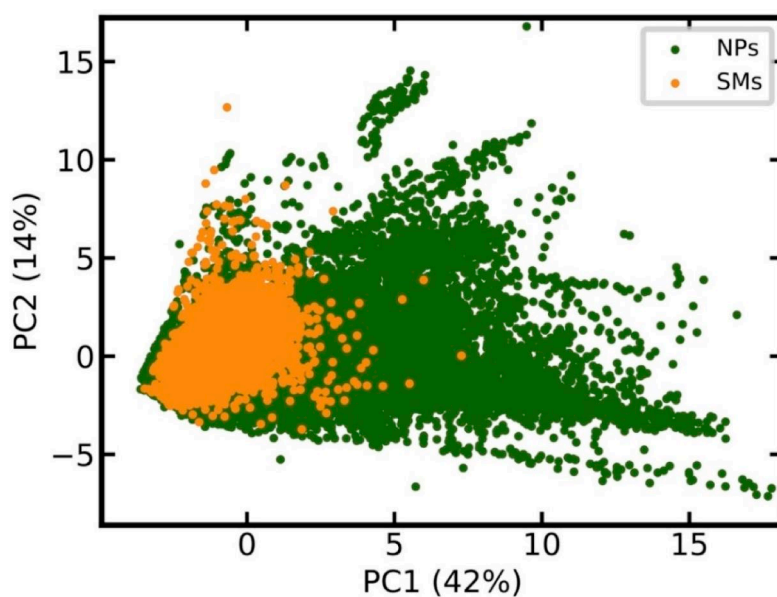


Figure 1. Comparison of the chemical space covered by natural products (NPs) and synthetic organic molecules (SMs). The score plot is based on the principle component analysis (PCA) of all molecules in the data set, characterized by 15 calculated physicochemical properties. PCA was performed on the full data sets. For the sake of clarity, only a randomly selected 10% of all data points are reported in the score plot. The percentage of the total variance explained by the first two principal components is reported in the respective axis labels.

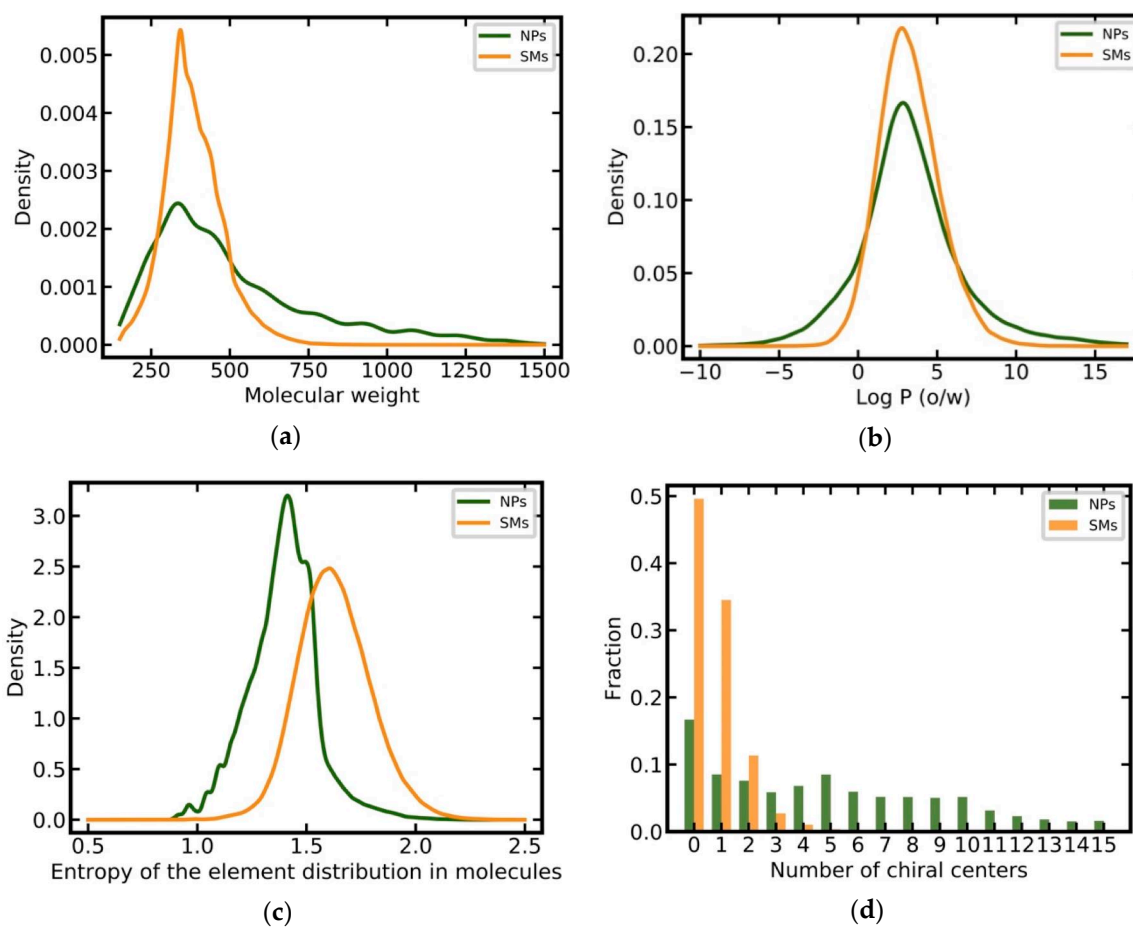


Figure 2. Cont.

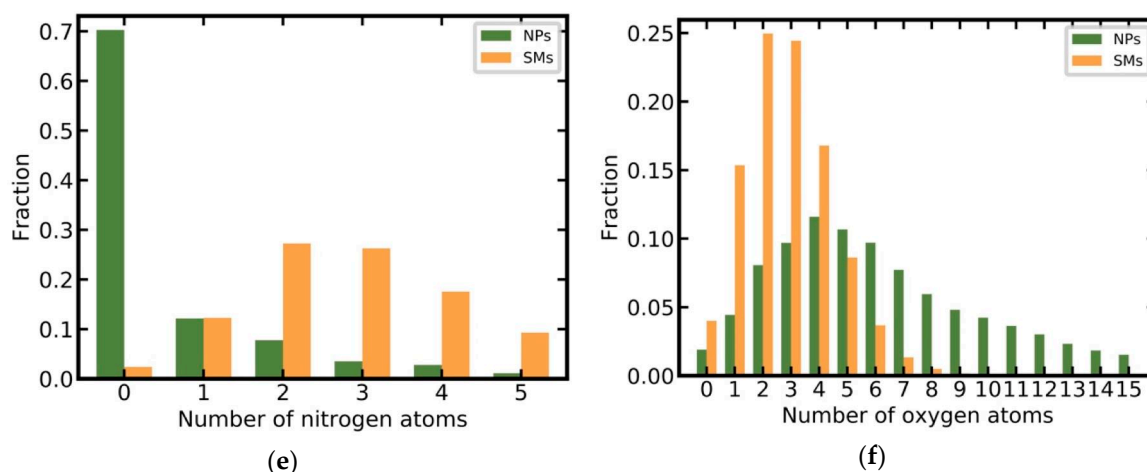


Figure 2. Distributions of key physicochemical properties among NPs and SMs: (a) Molecular weight; (b) $\log P$ (o/w); (c) entropy of the element distribution in molecules; (d) number of chiral centers; (e) number of nitrogen atoms; (f) number of oxygen atoms.

3.3. Model Development and Selection

Random forest classifiers [98] were trained on three different descriptor sets: 206 two-dimensional physicochemical property descriptors calculated with MOE [47], Morgan2 fingerprints (1024 bits) [48,49] calculated with RDKit [37], and MACCS keys (166 bits), also calculated with RDKit. Model performance was characterized utilizing the Matthews correlation coefficient (MCC) [99] and area under the receiver operating characteristic curve (AUC). The MCC is one of the most robust measures for evaluating the performance of binary classifiers, as it considers the proportion of all classes in the confusion matrix (i.e., true positives, false positives, true negatives, and false negatives). The AUC was used to measure how well the models are able to rank NPs early in a list.

As reported in Table 2, the models derived from any of the three descriptor sets performed very well. The AUC values, that were obtained during 10-fold cross-validation, were between 0.996 and 0.997; the MCC values were 0.950 or higher. No noticeable increase in performance was obtained by the further increase in the number of estimators ($n_{\text{estimators}}$) and the optimization of the maximum fraction of features considered per split (max_features ; data not shown). Therefore, we chose to use 100 estimators, and the square root of the number of features, as the most suitable setup for model generation.

Table 2. Performance of models derived from different descriptors or fingerprints.

Test Method	Metric ¹	MOE Two-Dimensional Descriptors	Morgan2 Fingerprints (1024 Bits)	MACCS Keys	NP-Likeness Calculator
10-fold cross-validation	AUC	0.997	0.997	0.996	/
	MCC	0.953	0.958	0.950	/
Independent test set	AUC	0.997	0.997	0.997	0.997
	MCC	0.954	0.960	0.960	0.959

¹ AUC: area under the receiver operating characteristic curve; MCC: Matthews correlation coefficient.

3.4. Model Validation

In a first step, the performance of the selected models was tested on an independent test set. The AUC and MCC values, that were obtained for the selected models on this independent test set, are comparable with those obtained for the 10-fold cross-validation: AUC values were 0.997 for models based on any of the three types of descriptors and MCC values were 0.954 or higher.

Given the fact that the type of descriptor, used for model generation, did not have a substantial impact on model performance, we opted to select the model based on MACCS keys as the primary model for further experiments, because of its low complexity and good interpretability. This model achieved a very good separation of NPs and SMs for the independent test set, as shown in Figure 3a. Approximately 63% of all NPs were assigned an NP class probability of 1.0, whereas 51% SMs were assigned an NP class probability of 0.0. Only approximately 1% of all compounds were assigned values close to the decision threshold of 0.5 (i.e., between 0.4 and 0.6).

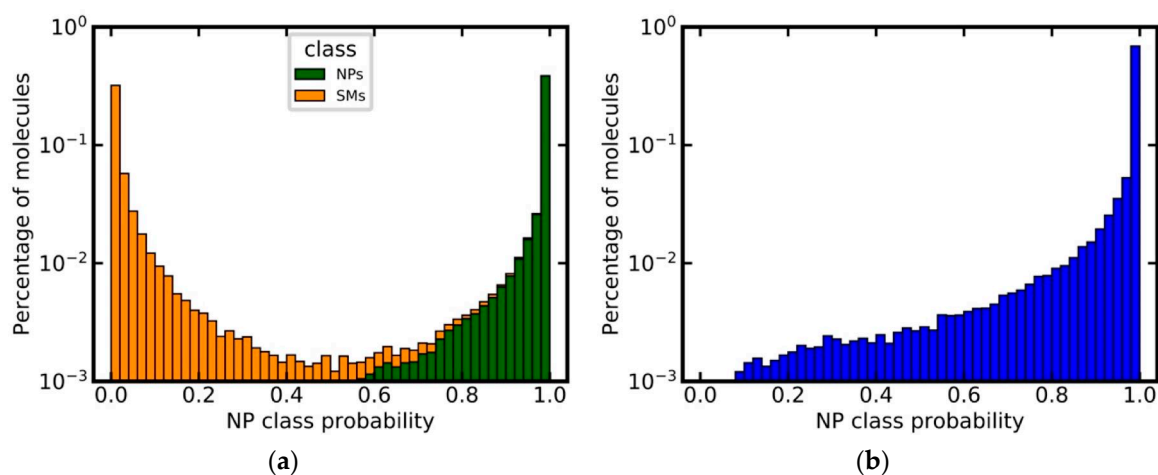


Figure 3. Predicted NP class probabilities distributions for (a) the independent test set (stacked histogram), (b) the DNP (after the removal of any compounds present in the training set). Note that the y-axis is in logarithmic scale.

The model's ability to identify NPs was also tested using the DNP as an external validation set. By definition, the DNP should consist exclusively of NPs. After the removal of any molecules present in the training data (based on canonicalized SMILES), the preprocessed DNP consisted of 60,502 compounds. Approximately 95% of these compounds were predicted as NPs by the model, demonstrating the model's capacity to identify NPs with high sensitivity (Figure 3b).

3.5. Comparison of Model Performance with the NP-Likeness Calculator

We compared the performance of the model derived from MACCS keys to the NP-likeness calculator (based on the Natural-Product-Likeness Scoring System; see Introduction), which we trained and tested on the identical data sets used for the development of our models. On the independent test set, the NP-likeness calculator performed equally well as our model, with an AUC of 0.997 and an MCC of 0.959 (Table 2). Approximately 95% of all compounds of the DNP were classified as NPs (i.e., having assigned an NP-likeness score greater than 0; see Figure S1), which is comparable to the classification obtained with our model based on MACCS keys.

3.6. Analysis of Class Probability Distributions for Different Data Sets

In addition to the above experiments, we used the model based on MACCS keys for profiling the ChEMBL database and a subset thereof. The ChEMBL database [44] primarily contains SMs, and 87% of all compounds stored in ChEMBL were predicted as such (Figure 4a). Interestingly, 42,949 molecules (~3%) were assigned an NP class probability of 1.0, and therefore likely are NPs. This finding is in agreement with our previous study, which identified approximately 40,000 NPs in the ChEMBL database, by overlapping the database with a comprehensive set of known NPs [19].

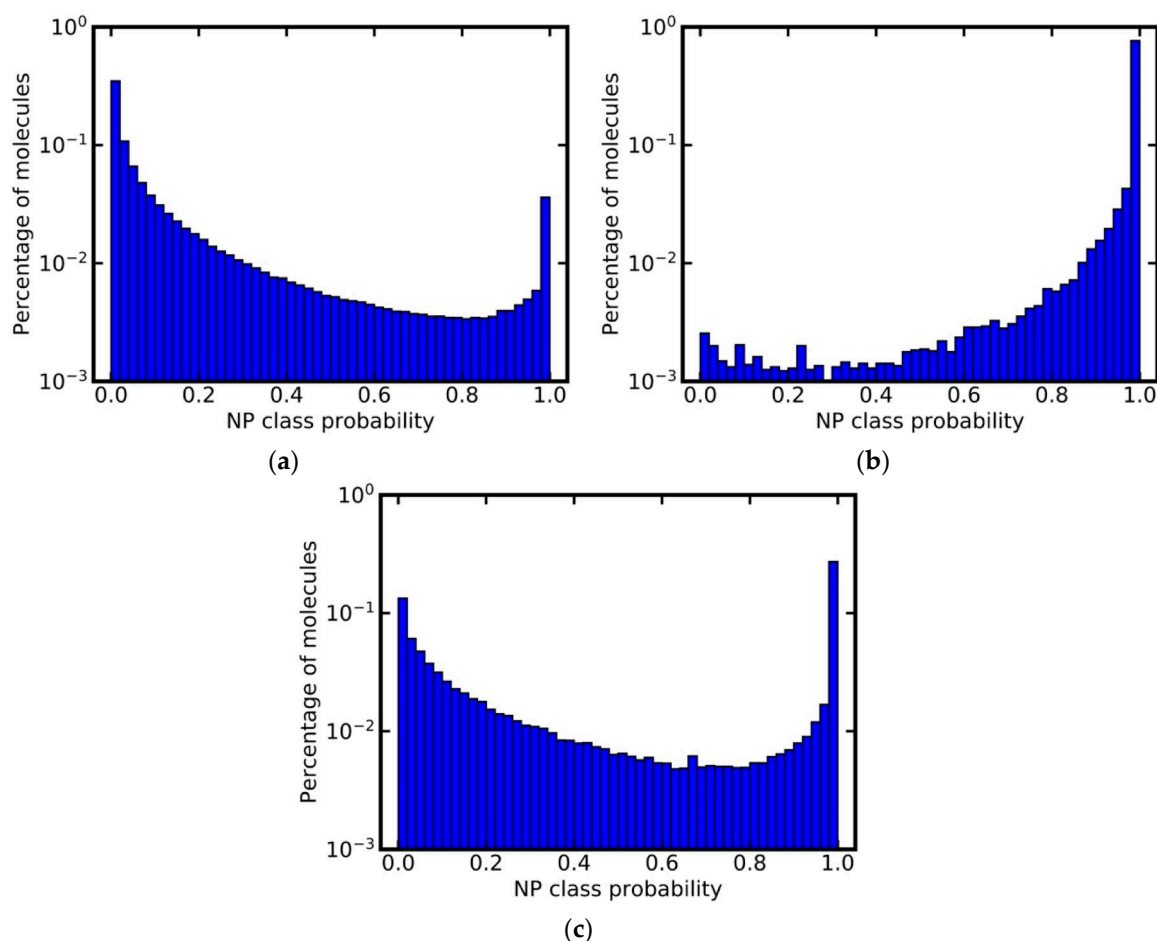


Figure 4. Predicted NP class probability distributions for (a) the ChEMBL database, (b) a subset of the ChEMBL database composed of molecules originating from the Journal of Natural Products, and (c) the natural products subset of ZINC. Note that the y-axis is in logarithmic scale.

A subset of the ChEMBL database containing molecules originating from the Journal of Natural Products [45] has been used as a source of genuine NPs to train models for the prediction of NP-likeness [31]. Our model based on MACCS keys predicts a small percentage of the molecules (less than 4%) in this data set as not NP-like (Figure 4b). Closer inspection of the compounds predicted as not NP-like reveals that these are, for example, SMs used as positive controls in biochemical assays. They include the drugs celecoxib, glibenclamide and linezolid, all of which are predicted with an NP class probability of 0.0. This experiment demonstrates that the classifiers can be used as powerful tools for the identification of NPs or SMs in mixed data sets with high accuracy.

A second example of a data set that by its name is assumed to consist exclusively of NPs is the natural products subset of ZINC [46]. The class probability distribution calculated for this subset however is similar to that obtained for the complete ChEMBL, indicating the presence of a substantial number of SMs (including NP derivatives and NP analogs) in this subset (Figure 4c): Only approximately 43% of all compounds in the NPs subset of ZINC were classified as NPs; around 23% were assigned an NP class probability of 1.0.

3.7. Analysis of Discriminative Features of Natural Products and Synthetic Molecules

The most discriminative features were determined, based on the `feature_importances_` attributes computed with scikit-learn (see Methods for details). For the classifier based on MOE two-dimensional molecular descriptors, the three most important features were the number of nitrogen atoms (a large fraction of NPs has no nitrogen atom; see Figure 2e), the entropy of the element distribution in molecules (NPs have on average lower element distribution entropy than SMs; see Figure 2c), and the number of unconstrained chiral centers (NPs have on average more chiral centers than SMs; see Figure 2d). An overview of the ten most important features is provided in Table 3.

Table 3. Feature importance for the random forest classifier based on MOE two-dimensional descriptors.

Identifier Used by MOE	Feature Importance ¹	Description
a_nN	0.103	Number of nitrogen atoms.
a_ICM	0.051	Entropy of the element distribution in the molecule.
chiral_u	0.045	Number of unconstrained chiral centers.
GCUT_SLOGP_0	0.045	Descriptor derived from graph distance adjacency matrices utilizing atomic contribution to log <i>P</i> .
SlogP_VSA0	0.044	Surface area descriptor taking into account the contributions of individual atoms to log <i>P</i> .
chiral	0.042	Number of chiral centers.
GCUT_SLOGP_3	0.036	Descriptor derived from graph distance adjacency matrices utilizing atomic contribution to log <i>P</i> .
a_nO	0.025	The number of oxygen atoms.
GCUT_PEOE_0	0.025	Descriptor derived from graph distance adjacency matrices utilizing partial equalization of orbital electronegativities charges.
SlogP_VSA1	0.024	Surface area descriptor taking into account the contributions of individual atoms to log <i>P</i> .

¹ From the `feature_importances_` attribute of the classifier based on MOE two-dimensional descriptors. The higher, the more important the feature is.

For the classifier based on MACCS keys, the 15 most important features are reported in Figure 5. In agreement with the differences observed in the physicochemical property distributions of NPs versus SMs (see Analysis of the Physicochemical Properties of Natural Products and Synthetic Molecules), the most important MACCS keys describe the presence or absence of nitrogen atoms, such as key 161, matching molecules containing at least one nitrogen atom, key 142, matching molecules with at least two nitrogen atoms, and keys 117, 158, 122, 156, 75, 110, 133, 92 and 80, matching molecules containing specific nitrogen-containing substructures. Also several oxygen-containing substructures are among the most important features, such as keys 139, 117, 110, 92.

vorapaxar shows that the model correctly identifies the decahydronaphtho[2,3-c]furan-1(3H)-one as NP-like, whereas it associates the modified areas with synthetic molecules. In the case of empagliflozin, which mimics the flavonoid phlorozin, the model correctly recognizes the C-glycosyl moiety as NP-like, whereas other atoms in the molecule are associated with synthetic molecules.

Table 4. Examples of similarity maps generated by the NP classifier based on Morgan2 fingerprints.

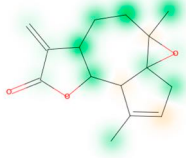
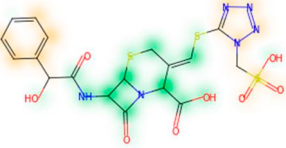
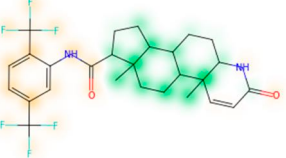
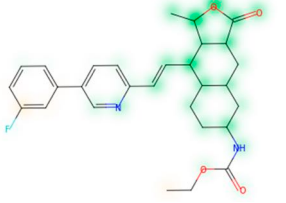
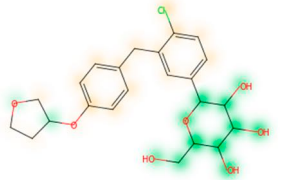
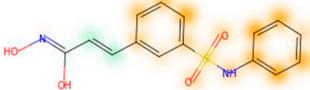
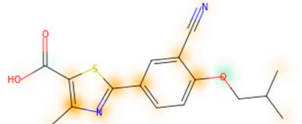
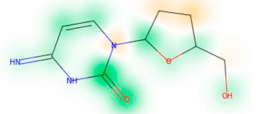
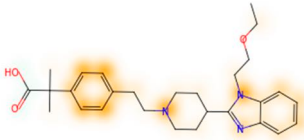
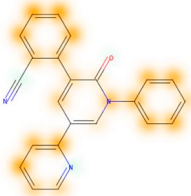
Similarity Map ¹	Name	Source ²	NP Class Probability	Disease Indication	Year Introduced
	arglabin	N	1.0	anticancer	1999
	cefonicid sodium	ND	0.34	antibacterial	1984
	dutasteride	ND	0.18	benign prostatic hypertrophy	2001
	vorapaxar	ND	0.30	coronary artery disease	2014
	empagliflozin	S*/NM	0.67	antidiabetic (diabetes 2)	2014
	belinostat	S*/NM	0.09	anticancer	2014
	febuxostat	S/NM	0.19	hyperuricemia	2009
	zalcitabine	S*	0.46	antiviral	1992

Table 4. Cont.

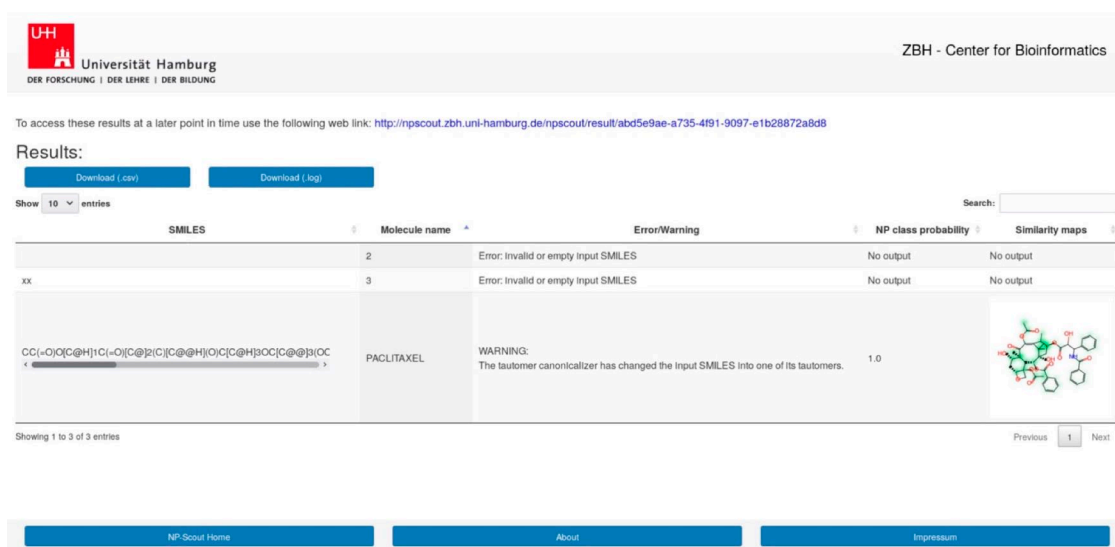
Similarity Map ¹	Name	Source ²	NP Class Probability	Disease Indication	Year Introduced
	bilastine	S	0.17	antihistamine	2011
	perampanel	S	0.16	antiepileptic	2012

¹ Green highlights mark atoms contributing to the classification of a molecule as NP, whereas orange highlights mark atoms contributing to the classification of a molecule as SM. ² N: Unaltered NP; ND: NP derivative; S*: Synthetic drug (NP pharmacophore); S: Synthetic drug; NM: Mimic of NP. Definitions according to ref [5].

3.9. NP-Scout Web Service

A web service named “NP-Scout” is accessible free of charge via <http://npscout.zbh.uni-hamburg.de/npscout>. It features the random forest model, based on MACCS keys for the computation of NP class probabilities and the random forest model, based on Morgan2 fingerprints (with 1024 bits) for the generation of similarity maps.

Users can submit molecular structures for calculation, by entering SMILES, uploading a file with SMILES or a list of SMILES, or drawing the molecule with the JavaScript Molecule Editor (JSME) [102]. The results page (Figure 6) presents the calculated NP class probabilities and similarity maps of submitted molecules in a tabular format. The results can be downloaded in CSV file format. Calculations of the NP class probabilities and the similarity maps take few seconds per compound and approximately 15 min for 1000 compounds. Users may utilize a unique link provided upon job submission to return to the website after all calculations have been completed.



U+H Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

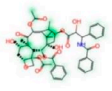
ZBH - Center for Bioinformatics

To access these results at a later point in time use the following web link: <http://npscout.zbh.uni-hamburg.de/npscout/result/abd5e9ae-a735-4f91-9097-e1b28872a8d8>

Results:

Download (.csv) Download (.log)

Show 10 entries

SMILES	Molecule name	Error/Warning	NP class probability	Similarity maps
	2	Error: Invalid or empty input SMILES	No output	No output
xx	3	Error: Invalid or empty input SMILES	No output	No output
<chem>CC(=O)OC@H]1C(-O)[C@]2[C][C@@H]O[C]C@H]3OC[C@@]3[OC</chem>	PACLITAXEL	WARNING: The tautomer canonicalizer has changed the input SMILES into one of its tautomers.	1.0	

Showing 1 to 3 of 3 entries

Previous 1 Next

NP-Scout Home About Impressum

Figure 6. Screenshot of the result page of NP-Scout.

4. Conclusions

In this work, we introduced a pragmatic machine learning approach for the discrimination of NPs and SMs and for the quantification of NP-likeness. As shown by validation experiments using independent and external testing data, the models reach a very high level of accuracy. An interesting and relevant new aspect of this work is the utilization of similarity maps to visualize atoms in molecules making decisive contributions to the assignment of compounds to either class. A free web service for the classification of small molecules and the visualization of similarity maps is available at <http://npscout.zbh.uni-hamburg.de/npscout>.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2218-273X/9/2/43/s1>, Figure S1: Distribution of calculated NP-likeness scores for the DNP (after removal of any compounds present in the training set).

Author Contributions: Conceptualization, Y.C. and J.K.; methodology, Y.C. and J.K.; software, Y.C., C.S., and S.H.; validation, Y.C.; formal analysis, Y.C.; investigation, Y.C., C.S., and S.H.; resources, J.K.; data curation, Y.C.; writing—original draft preparation, Y.C., C.S., S.H., and J.K.; visualization, Y.C. and S.H.; supervision, J.K.; project administration, J.K.; funding acquisition, Y.C. and J.K.

Funding: Y.C. is supported by the China Scholarship Council, grant number 201606010345. C.S. and J.K. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number KI 2085/1-1. J.K. is also supported by the Bergens Forskningsstiftelse (BFS, Bergen Research Foundation), grant number BFS2017TMT01.

Acknowledgments: Gerd Embruch from the Center of Bioinformatics (ZBH) of the Universität Hamburg is thanked for his technical support with the web service.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Cragg, G.M.; Newman, D.J. Biodiversity: A continuing source of novel drug leads. *J. Macromol. Sci. Part A Pure Appl. Chem.* **2005**, *77*, 7–24. [[CrossRef](#)]
2. Rodrigues, T.; Reker, D.; Schneider, P.; Schneider, G. Counting on natural products for drug design. *Nat. Chem.* **2016**, *8*, 531–541. [[CrossRef](#)] [[PubMed](#)]
3. Harvey, A.L.; Edrada-Ebel, R.; Quinn, R.J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **2015**, *14*, 111–129. [[CrossRef](#)]
4. Shen, B. A new golden age of natural products drug discovery. *Cell* **2015**, *163*, 1297–1300. [[CrossRef](#)] [[PubMed](#)]
5. Newman, D.J.; Cragg, G.M. Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79*, 629–661. [[CrossRef](#)] [[PubMed](#)]
6. Grabowski, K.; Baringhaus, K.-H.; Schneider, G. Scaffold diversity of natural products: Inspiration for combinatorial library design. *Nat. Prod. Rep.* **2008**, *25*, 892–904. [[CrossRef](#)]
7. Ertl, P.; Schuffenhauer, A. Cheminformatics analysis of natural products: Lessons from nature inspiring the design of new drugs. *Prog. Drug Res.* **2008**, *66*, 219–235.
8. Chen, Y.; de Lomana, M.G.; Friedrich, N.-O.; Kirchmair, J. Characterization of the chemical space of known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58*, 1518–1532. [[CrossRef](#)]
9. Chen, H.; Engkvist, O.; Blomberg, N.; Li, J. A comparative analysis of the molecular topologies for drugs, clinical candidates, natural products, human metabolites and general bioactive compounds. *Med. Chem. Commun.* **2012**, *3*, 312–321. [[CrossRef](#)]
10. Camp, D.; Gavelas, A.; Campitelli, M. Analysis of physicochemical properties for drugs of natural origin. *J. Nat. Prod.* **2015**, *78*, 1370–1382. [[CrossRef](#)]
11. Koch, M.A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17272–17277. [[CrossRef](#)] [[PubMed](#)]
12. Stratton, C.F.; Newman, D.J.; Tan, D.S. Cheminformatic comparison of approved drugs from natural product versus synthetic origins. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 4802–4807. [[CrossRef](#)] [[PubMed](#)]

13. Wetzel, S.; Schuffenhauer, A.; Roggo, S.; Ertl, P.; Waldmann, H. Cheminformatic analysis of natural products and their chemical space. *CHIMIA Int. J. Chem.* **2007**, *61*, 355–360. [[CrossRef](#)]
14. López-Vallejo, F.; Giulianotti, M.A.; Houghten, R.A.; Medina-Franco, J.L. Expanding the medically relevant chemical space with compound libraries. *Drug Discov. Today* **2012**, *17*, 718–726. [[CrossRef](#)]
15. Feher, M.; Schmidt, J.M. Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227. [[CrossRef](#)] [[PubMed](#)]
16. Clemons, P.A.; Bodycombe, N.E.; Carrinski, H.A.; Wilson, J.A.; Shamji, A.F.; Wagner, B.K.; Koehler, A.N.; Schreiber, S.L. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 18787–18792. [[CrossRef](#)] [[PubMed](#)]
17. Henkel, T.; Brunne, R.M.; Müller, H.; Reichel, F. Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angew. Chem. Int. Ed. Engl.* **1999**, *38*, 643–647. [[CrossRef](#)]
18. Lee, M.L.; Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: Application in the design of natural product-based combinatorial libraries. *J. Comb. Chem.* **2001**, *3*, 284–289. [[CrossRef](#)]
19. Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data resources for the computer-guided discovery of bioactive natural products. *J. Chem. Inf. Model.* **2017**, *57*, 2099–2111. [[CrossRef](#)]
20. Rupp, M.; Schroeter, T.; Steri, R.; Zettl, H.; Proschak, E.; Hansen, K.; Rau, O.; Schwarz, O.; Müller-Kuhrt, L.; Schubert-Zsilavecz, M.; et al. From machine learning to natural product derivatives that selectively activate transcription factor PPAR γ . *ChemMedChem* **2010**, *5*, 191–194. [[CrossRef](#)]
21. Maindola, P.; Jamal, S.; Grover, A. Cheminformatics based machine learning models for AMA1-RON2 abrogators for inhibiting Plasmodium falciparum erythrocyte invasion. *Mol. Inform.* **2015**, *34*, 655–664. [[CrossRef](#)] [[PubMed](#)]
22. Chagas-Paula, D.A.; Oliveira, T.B.; Zhang, T.; Edrada-Ebel, R.; Da Costa, F.B. Prediction of anti-inflammatory plants and discovery of their biomarkers by machine learning algorithms and metabolomic studies. *Planta Med.* **2015**, *81*, 450–458. [[CrossRef](#)] [[PubMed](#)]
23. Reker, D.; Perna, A.M.; Rodrigues, T.; Schneider, P.; Reutlinger, M.; Mönch, B.; Koeberle, A.; Lamers, C.; Gabler, M.; Steinmetz, H.; et al. Revealing the macromolecular targets of complex natural products. *Nat. Chem.* **2014**, *6*, 1072–1078. [[CrossRef](#)] [[PubMed](#)]
24. Rodrigues, T.; Sieglitz, F.; Somovilla, V.J.; Cal, P.M.S.D.; Galione, A.; Corzana, F.; Bernardes, G.J.L. Unveiling (–)-englerin A as a modulator of L-type calcium channels. *Angew. Chem. Int. Ed. Engl.* **2016**, *55*, 11077–11081. [[CrossRef](#)] [[PubMed](#)]
25. Merk, D.; Grisoni, F.; Friedrich, L.; Gelzinyte, E.; Schneider, G. Computer-assisted discovery of retinoid X receptor modulating natural products and isofunctional mimetics. *J. Med. Chem.* **2018**, *61*, 5442–5447. [[CrossRef](#)] [[PubMed](#)]
26. Schneider, P.; Schneider, G. De-orphaning the marine natural product (\pm)-marinopyrrole A by computational target prediction and biochemical validation. *Chem. Commun.* **2017**, *53*, 2272–2274.
27. Merk, D.; Grisoni, F.; Friedrich, L.; Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Commun. Chem.* **2018**, *1*, 68.
28. Friedrich, L.; Rodrigues, T.; Neuhaus, C.S.; Schneider, P.; Schneider, G. From complex natural products to simple synthetic mimetics by computational de novo design. *Angew. Chem. Int. Ed. Engl.* **2016**, *55*, 6789–6792. [[CrossRef](#)]
29. Grisoni, F.; Merk, D.; Consonni, V.; Hiss, J.A.; Tagliabue, S.G.; Todeschini, R.; Schneider, G. Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Commun. Chem.* **2018**, *1*, 44.
30. Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.* **2008**, *48*, 68–74. [[CrossRef](#)]
31. Jayaseelan, K.V.; Moreno, P.; Truszkowski, A.; Ertl, P.; Steinbeck, C. Natural product-likeness score revisited: An open-source, open-data implementation. *BMC Bioinform.* **2012**, *13*, 106. [[CrossRef](#)] [[PubMed](#)]
32. Jayaseelan, K.V.; Steinbeck, C. Building blocks for automated elucidation of metabolites: Natural product-likeness for candidate ranking. *BMC Bioinform.* **2014**, *15*, 234. [[CrossRef](#)] [[PubMed](#)]
33. RDKit NP_Score. Available online: https://github.com/rdkit/rdkit/tree/master/Contrib/NP_Score (accessed on 27 November 2018).

34. Yu, M.J. Natural product-like virtual libraries: Recursive atom-based enumeration. *J. Chem. Inf. Model.* **2011**, *51*, 541–557. [[CrossRef](#)] [[PubMed](#)]
35. Zaid, H.; Raiyn, J.; Nasser, A.; Saad, B.; Rayan, A. Physicochemical properties of natural based products versus synthetic chemicals. *Open Nutraceuticals J.* **2010**, *3*, 194–202. [[CrossRef](#)]
36. Riniker, S.; Landrum, G.A. Similarity maps—A visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminform.* **2013**, *5*, 43. [[CrossRef](#)] [[PubMed](#)]
37. RDKit Version 2017.09.3: Open-source cheminformatics software. Available online: <http://www.rdkit.org> (accessed on 22 May 2018).
38. Stork, C.; Wagner, J.; Friedrich, N.-O.; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: A machine-learning model for the prediction of frequent hitters. *ChemMedChem* **2018**, *13*, 564–571. [[CrossRef](#)]
39. MolVS Version 0.1.1. Available online: <https://github.com/mcs07/MolVS> (accessed on 12 July 2018).
40. Sterling, T.; Irwin, J.J. ZINC 15-Ligand discovery for everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. [[CrossRef](#)]
41. ZINC “in-stock” subset. ZINC15. Available online: <http://zinc15.docking.org/> (accessed on 21 August 2018).
42. *Dictionary of Natural Products*, version 19.1; Chapman & Hall/CRC: London, UK, 2010.
43. Bento, A.P.; Gaulton, A.; Hersey, A.; Bellis, L.J.; Chambers, J.; Davies, M.; Krüger, F.A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090. [[CrossRef](#)]
44. ChEMBL Version 24_1. Available online: <https://www.ebi.ac.uk/chembl/> (accessed on 30 July 2018).
45. ChEMBL Version 23. Available online: <https://www.ebi.ac.uk/chembl/> (accessed on 6 June 2017).
46. Natural products subset of ZINC. ZINC15. Available online: <http://zinc15.docking.org/substances/subsets/> (accessed on 7 November 2018).
47. *Molecular Operating Environment (MOE)*, version 2016.08; Chemical Computing Group: Montreal, QC, Canada, 2016.
48. Morgan, H.L. The generation of a unique machine description for chemical structures—A technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113. [[CrossRef](#)]
49. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [[CrossRef](#)]
50. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
51. Scikit-Learn: Machine Learning in Python. version 0.19.1.
52. Natural Product Likeness Calculator Version 2.1. Available online: <https://sourceforge.net/projects/np-likeness/> (accessed on 5 October 2018).
53. Natural Products Atlas. Available online: <https://www.npatlas.org/> (accessed on 20 August 2018).
54. Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS ONE* **2013**, *8*, e62839. [[CrossRef](#)]
55. Universal Natural Products Database (UNPD). Available online: <http://pkuxj.pku.edu.cn/UNPD> (accessed on 17 October 2016).
56. Chen, C.Y.-C. TCM Database@Taiwan: The world’s largest traditional Chinese medicine database for drug screening in silico. *PLoS ONE* **2011**, *6*, e15939. [[CrossRef](#)] [[PubMed](#)]
57. TCM Database@Taiwan. Available online: <http://tcm.cmu.edu.tw> (accessed on 17 October 2016).
58. Xue, R.; Fang, Z.; Zhang, M.; Yi, Z.; Wen, C.; Shi, T. TCMID: Traditional Chinese medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Res.* **2013**, *41*, D1089–D1095. [[CrossRef](#)] [[PubMed](#)]
59. Traditional Chinese Medicine Integrated Database (TCMID). Available online: www.megabionet.org/tcmid (accessed on 19 October 2016).
60. Lin, Y.-C.; Wang, C.-C.; Chen, I.-S.; Jheng, J.-L.; Li, J.-H.; Tung, C.-W. TIPdb: A database of anticancer, antiplatelet, and antituberculosis phytochemicals from indigenous plants in Taiwan. *Sci. World J.* **2013**, *2013*, 736386. [[CrossRef](#)]
61. Tung, C.-W.; Lin, Y.-C.; Chang, H.-S.; Wang, C.-C.; Chen, I.-S.; Jheng, J.-L.; Li, J.-H. TIPdb-3D: The three-dimensional structure database of phytochemicals from Taiwan indigenous plants. *Database* **2014**, *2014*, bau055. [[CrossRef](#)] [[PubMed](#)]

62. Taiwan Indigenous Plant Database (TIPdb). Available online: <http://cwtung.kmu.edu.tw/tipdb> (accessed on 19 October 2016).
63. Ambinter. Available online: www.ambinter.com (accessed on 2 June 2017).
64. GreenPharma. Available online: www.greenpharma.com (accessed on 2 June 2017).
65. AnalytiCon Discovery. Available online: www.ac-discovery.com (accessed on 14 November 2017).
66. Ntie-Kang, F.; Telukunta, K.K.; Döring, K.; Simoben, C.V.; A Moumbock, A.F.; Malange, Y.I.; Njume, L.E.; Yong, J.N.; Sippl, W.; Günther, S. NANPDB: A resource for natural products from Northern African sources. *J. Nat. Prod.* **2017**, *80*, 2067–2076. [[CrossRef](#)]
67. Northern African Natural Products Database (NANPDB). Available online: www.african-compounds.org/nanpdb (accessed on 5 April 2017).
68. Klementz, D.; Döring, K.; Lucas, X.; Telukunta, K.K.; Erxleben, A.; Deubel, D.; Erber, A.; Santillana, I.; Thomas, O.S.; Bechthold, A.; et al. StreptomeDB 2.0—An extended resource of natural products produced by streptomycetes. *Nucleic Acids Res.* **2015**, *44*, D509–D514. [[CrossRef](#)]
69. StreptomeDB. Available online: <http://132.230.56.4/streptomedb2/> (accessed on 13 April 2017).
70. Ming, H.; Tiejun, C.; Yanli, W.; Stephen, B.H. Web search and data mining of natural products and their bioactivities in PubChem. *Sci. China Chem.* **2013**, *56*, 1424–1435.
71. Natural products subset. PubChem Substance Database. Available online: <http://ncbi.nlm.nih.gov/pcsubstance> (accessed on 7 April 2017).
72. Pilon, A.C.; Valli, M.; Dametto, A.C.; Pinto, M.E.F.; Freire, R.T.; Castro-Gamboa, I.; Andricopulo, A.D.; Bolzani, V.S. NuBBE: An updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci. Rep.* **2017**, *7*, 7215. [[CrossRef](#)]
73. Núcleo de Bioensaios, Biossíntese e Ecofisiologia de Produtos Naturais (NuBBE). Available online: <http://nubbe.iq.unesp.br/portal/nubbedb.html> (accessed on 19 April 2017).
74. PI Chemicals. Available online: www.pipharm.com (accessed on 5 May 2017).
75. Choi, H.; Cho, S.Y.; Pak, H.J.; Kim, Y.; Choi, J.-Y.; Lee, Y.J.; Gong, B.H.; Kang, Y.S.; Han, T.; Choi, G.; et al. NPCARE: Database of natural products and fractional extracts for cancer regulation. *J. Cheminform.* **2017**, *9*, 2. [[CrossRef](#)] [[PubMed](#)]
76. Database of Natural Products for Cancer Gene Regulation (NPCARE). Available online: <http://silver.sejong.ac.kr/npcare> (accessed on 20 February 2017).
77. Mangal, M.; Sagar, P.; Singh, H.; Raghava, G.P.S.; Agarwal, S.M. NPACT: Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database. *Nucleic Acids Res.* **2013**, *41*, D1124–D1129. [[CrossRef](#)] [[PubMed](#)]
78. Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database (NPACT). Available online: <http://crdd.osdd.net/raghava/npact> (accessed on 13 April 2017).
79. InterBioScreen. Available online: www.ibscreen.com (accessed on 14 November 2017).
80. Ntie-Kang, F.; Zofou, D.; Babiaka, S.B.; Meudom, R.; Scharfe, M.; Lifongo, L.L.; Mbah, J.A.; Mbaze, L.M.; Sippl, W.; Efang, S.M.N. AfroDb: A select highly potent and diverse natural product library from African medicinal plants. *PLoS ONE* **2013**, *8*, e78085. [[CrossRef](#)] [[PubMed](#)]
81. AfroDb. Available online: <http://african-compounds.org/about/afrodb> (accessed on 18 October 2016).
82. TargetMol. Available online: www.targetmol.com (accessed on 17 May 2017).
83. Kang, H.; Tang, K.; Liu, Q.; Sun, Y.; Huang, Q.; Zhu, R.; Gao, J.; Zhang, D.; Huang, C.; Cao, Z. HIM-herbal ingredients in-vivo metabolism database. *J. Cheminform.* **2013**, *5*, 28. [[CrossRef](#)] [[PubMed](#)]
84. Herbal Ingredients In-Vivo Metabolism database (HIM). Available online: <http://binfo.shmtu.edu.cn:8080/him> (accessed on 13 April 2017).
85. Hatherley, R.; Brown, D.K.; Musyoka, T.M.; Penkler, D.L.; Faya, N.; Lobb, K.A.; Tastan Bishop, Ö. SANCDB: A South African natural compound database. *J. Cheminform.* **2015**, *7*, 29. [[CrossRef](#)] [[PubMed](#)]
86. South African Natural Compound Database (SANCDB). Available online: <http://sancdb.rubi.ru.ac.za> (accessed on 8 February 2017).
87. UEFS Natural Products Catalog. ZINC15. Available online: <http://zinc15.docking.org> (accessed on 26 May 2017).
88. Ntie-Kang, F.; Amoa Onguéné, P.; Fotso, G.W.; Andrae-Marobela, K.; Bezabih, M.; Ndom, J.C.; Ngadjui, B.T.; Ogundaini, A.O.; Abegaz, B.M.; Meva'a, L.M. Virtualizing the p-ANAPL library: A step towards drug discovery from African medicinal plants. *PLoS ONE* **2014**, *9*, e90655. [[CrossRef](#)]

89. Natural Products Set IV of the Developmental Therapeutic Program of the National Cancer Institute/National Institutes of Health. Available online: http://dtp.cancer.gov/organization/dscb/obtaining/available_plates.htm (accessed on 20 October 2016).
90. Ye, H.; Ye, L.; Kang, H.; Zhang, D.; Tao, L.; Tang, K.; Liu, X.; Zhu, R.; Liu, Q.; Chen, Y.Z.; et al. HIT: Linking herbal active ingredients to targets. *Nucleic Acids Res.* **2011**, *39*, D1055–D1059. [CrossRef]
91. Herbal Ingredients' Targets database (HIT). Available online: <http://lifecenter.sgst.cn/hit> (accessed on 13 April 2017).
92. Ntie-Kang, F.; Nwodo, J.N.; Ibezim, A.; Simoben, C.V.; Karaman, B.; Ngwa, V.F.; Sippl, W.; Adikwu, M.U.; Mbaze, L.M. Molecular modeling of potential anticancer agents from African medicinal plants. *J. Chem. Inf. Model.* **2014**, *54*, 2433–2450. [CrossRef]
93. AfroCancer. Available online: <http://african-compounds.org/about/afrocancer> (accessed on 10 February 2017).
94. Onguéné, P.A.; Ntie-Kang, F.; Mbah, J.A.; Lifongo, L.L.; Ndom, J.C.; Sippl, W.; Mbaze, L.M. 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.* **2014**, *4*, 6. [CrossRef]
95. AfroMalariaDB. Available online: <http://african-compounds.org/about/afromalariadb> (accessed on 10 February 2017).
96. Natural products subset of AK Scientific. AK Scientific. Available online: www.aksci.com (accessed on 19 April 2017).
97. Natural products of Selleck Chemicals. Selleck Chemicals. Available online: www.selleckchem.com (accessed on 14 November 2017).
98. Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32. [CrossRef]
99. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451. [CrossRef]
100. Schomburg, K.; Ehrlich, H.-C.; Stierand, K.; Rarey, M. From structure diagrams to visual chemical patterns. *J. Chem. Inf. Model.* **2010**, *50*, 1529–1535. [CrossRef] [PubMed]
101. SMARTSview. Available online: <http://smartsview.zbh.uni-hamburg.de/> (accessed on 30 November 2018).
102. Bienfait, B.; Ertl, P. JSME: A free molecule editor in JavaScript. *J. Cheminform.* **2013**, *5*, 24. [CrossRef] [PubMed]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

3.3. Scope of 3D Shape-Based Approaches in Predicting the Macromolecular Targets of Structurally Complex Small Molecules

Knowing the target(s) of small molecules is one of the most important tasks to evaluate the pharmacological efficacy and safety of compounds, and for further optimization. However, even for many approved drugs their targets remain to be identified. There is an increasing number of computational methods for target prediction recently but predicting likely targets for natural products or in general for structurally complex small molecules (CSMs) is more challenging.

One challenge in target prediction for NPs is that there is much less bioactivity data compared to structurally less complex, synthetic molecules. NPs differ from synthetic molecules in terms of physicochemical and structural properties and many NPs have complex 3D molecular shapes. When using in silico methods trained on synthetic molecules to predict targets for NPs, extra caution should be exercised.

A 3D shape-based method was evaluated to predict the macromolecular targets of structurally CSMs, including natural products and macrocyclic ligands, and details of the study are shown in the following publication (D6).

[D6] Chen, Y.; Mathai, N.; Kirchmair, J. Scope of 3D Shape-Based Approaches in Predicting the Macromolecular Targets of Structurally Complex Small Molecules Including Natural Products and Macrocyclic Ligands. *J. Chem. Inf. Model.* **2020**, *60* (6), 2858-2875.

Available at <https://doi.org/10.1021/acs.jcim.0c00161>.

Y. Chen and J. Kirchmair conceptualized the work. Y. Chen collected, curated and analyzed all data. She also conducted all computational work. N. Mathai prepared and provided a high-quality data set from ChEMBL. Y. Chen wrote the largest part of the manuscript. J. Kirchmair supervised this work.

Reprinted from:

Chen, Y.; Mathai, N.; Kirchmair, J. Scope of 3D Shape-Based Approaches in Predicting the Macromolecular Targets of Structurally Complex Small Molecules Including Natural Products and Macrocyclic Ligands. *J. Chem. Inf. Model.* **2020**, *60* (6), 2858-2875.

This is an open access article published under a Creative Commons Attribution (CC-BY) License.

The supplementary information of this article can be found in [Appendix C](#).

Scope of 3D Shape-Based Approaches in Predicting the Macromolecular Targets of Structurally Complex Small Molecules Including Natural Products and Macrocyclic Ligands

Ya Chen, Neann Mathai, and Johannes Kirchmair*

Cite This: *J. Chem. Inf. Model.* 2020, 60, 2858–2875

Read Online

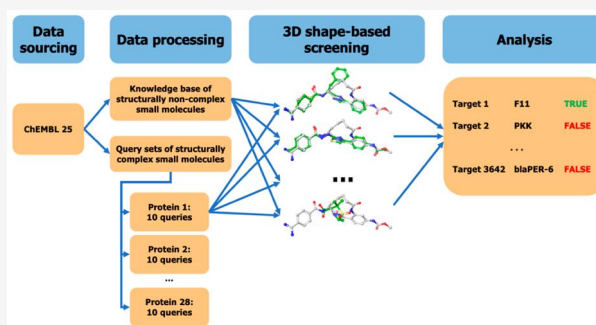
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: A plethora of similarity-based, network-based, machine learning, docking and hybrid approaches for predicting the macromolecular targets of small molecules are available today and recognized as valuable tools for providing guidance in early drug discovery. With the increasing maturity of target prediction methods, researchers have started to explore ways to expand their scope to more challenging molecules such as structurally complex natural products and macrocyclic small molecules. In this work, we systematically explore the capacity of an alignment-based approach to identify the targets of structurally complex small molecules (including large and flexible natural products and macrocyclic compounds) based on the similarity of their 3D molecular shape to noncomplex molecules (i.e., more conventional, “drug-like”, synthetic compounds). For this analysis, query sets of 10 representative, structurally complex molecules were compiled for each of the 28 pharmaceutically relevant proteins. Subsequently, ROCS, a leading shape-based screening engine, was utilized to generate rank-ordered lists of the potential targets of the 28 × 10 queries according to the similarity of their 3D molecular shapes with those of compounds from a knowledge base of 272 640 noncomplex small molecules active on a total of 3642 different proteins. Four of the scores implemented in ROCS were explored for target ranking, with the TanimotoCombo score consistently outperforming all others. The score successfully recovered the targets of 30% and 41% of the 280 queries among the top-5 and top-20 positions, respectively. For 24 out of the 28 investigated targets (86%), the method correctly assigned the first rank (out of 3642) to the target of interest for at least one of the 10 queries. The shape-based target prediction approach showed remarkable robustness, with good success rates obtained even for compounds that are clearly distinct from any of the ligands present in the knowledge base. However, complex natural products and macrocyclic compounds proved to be challenging even with this approach, although cases of complete failure were recorded only for a small number of targets.



INTRODUCTION

The past decade has seen a boost in the development of in silico approaches for the prediction of the macromolecular targets of small molecules.^{1–3} Progress has been fueled by, among other factors, (i) the increasing amount of chemical and biological data available in the public domain, (ii) the strategic shift from the “one drug-one target” paradigm that had dominated small-molecule drug discovery for decades to the concept of polypharmacology,⁴ and (iii) advances in computational power and algorithms. Despite the rapid development, however, it is challenging to obtain a realistic understanding of the performance of target prediction methods.⁵

There are several classes of in silico approaches for target prediction in existence: (i) similarity-based methods, which use the similarity between data such as small molecules, targets, and interactions to make predictions,⁶ (ii) network-based methods, where networks based on anything from ligand similarity⁷ to highly heterogeneous data are built to gain

systemic understanding of modeled data,⁸ (iii) machine learning approaches, which make use of machine learning methods such as random forests, support vector machines, or artificial neural networks to make predictions,⁹ (iv) reverse (or inverse) docking methods, which dock queries into potential targets to make predictions based on docking scores³ and methods which combine two or several types of these approaches.¹

A large proportion of models reported in the scientific literature are available as free public web services or commercial tools.¹⁰ Most models utilize information from

Received: February 13, 2020

Published: May 5, 2020



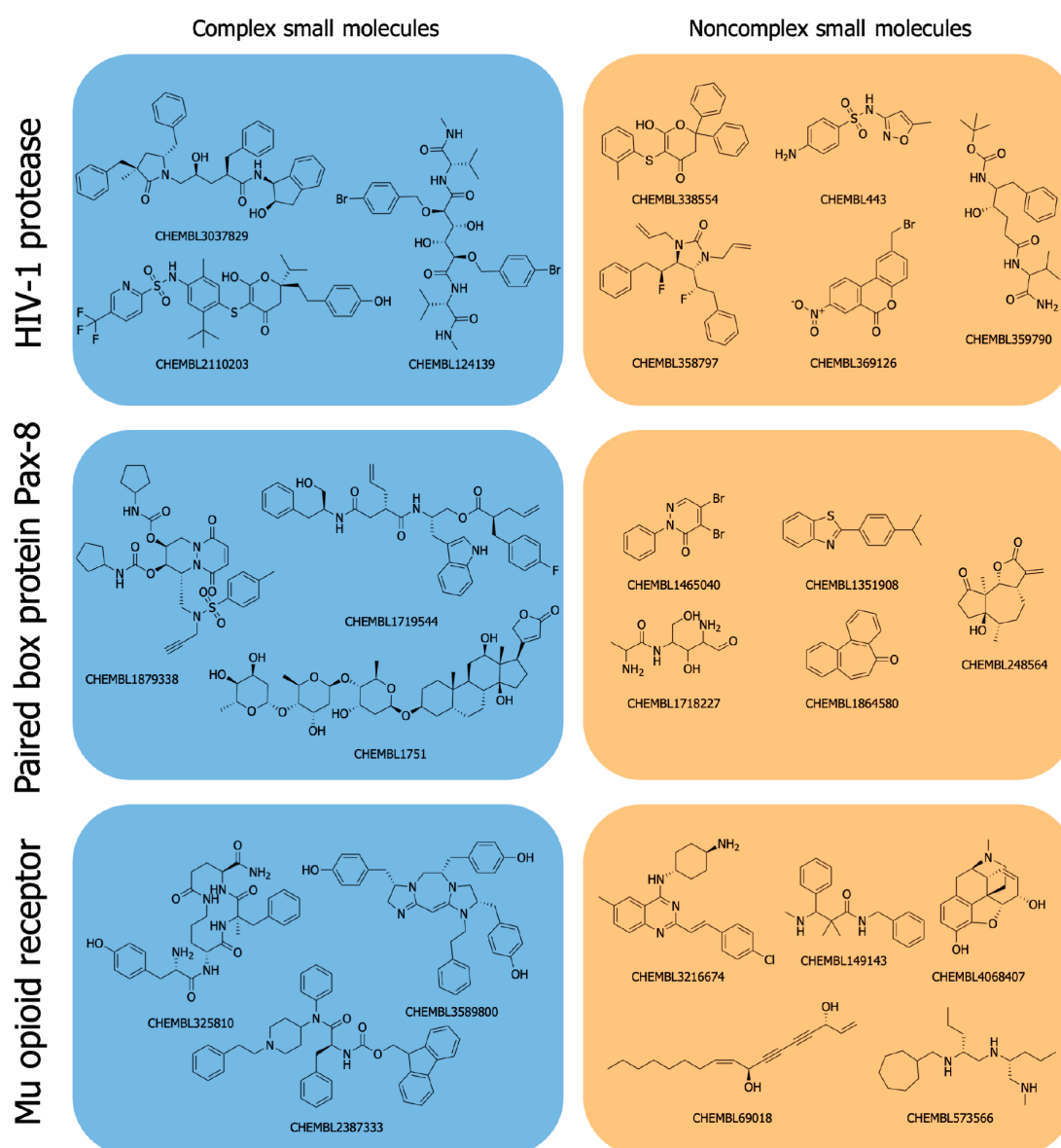


Figure 1. Examples of CSMs and non-CSMs. Represented on the left are the three most diverse CSMs (used as queries in this study) identified for the HIV-1 protease, paired box protein Pax-8 and mu opioid receptor, and on the right the five most diverse non-CSMs (representing the knowledge base compounds). More details on the automated and unbiased procedure employed for selecting these example compounds are provided in the [Compilation of a Test Set for Target Prediction](#) section in the [Methods](#) section.

the largest public resources of chemical and biological data, PubChem,¹¹ and the ChEMBL database.¹² PubChem currently contains more than 102 million compounds and 268 million bioactivity data points,¹³ and the latest release of the ChEMBL database contains close to 2 million compounds, with more than 16 million measured activities.¹⁴

With the increasing coverage and reliability of the models, researchers have started to develop strategies for predicting the likely targets of more challenging compounds such as natural products,^{15,16} for which there is a notorious lack of available measured data,¹⁷ and macrocyclic compounds, characterized by a large number of conformational degrees of freedom in combination with distinct torsional angle preferences.^{18–20} For example, Reker et al.²¹ dissected the macrocyclic antitumor agent archazolid A and used pharmacophoric descriptions of

these fragments to relate them to small molecules with known bioactivities. Several then unknown targets of archazolid A that were predicted by this approach have subsequently been confirmed in biological tests. More recently, Cockroft et al.¹⁶ have reported on the development of a stacked ensemble approach which, despite being trained on data for synthetic compounds, is able to predict the macromolecular targets of natural products with good accuracy.

In silico methods based on the comparison of the 3D molecular shapes of aligned molecules are predestined for use in target prediction because of their ability to recognize similarity among structurally dissimilar compounds, as long as their molecular shapes (or at least parts of their molecular shapes) are preserved. Most shape-based methods take the distribution of chemical features (“color”) into account, which

contributes substantially to their performance.²² They form the basis of several target prediction approaches^{23–25} and are also attractive tools for virtual screening and scaffold hopping.^{22,26,27}

Here, we systematically investigate the capacity of a leading 3D alignment-dependent, shape-based approach to identify the macromolecular targets of structurally complex small molecules (CSMs) on the basis of their molecular similarity with non-CSMs. In the context of small-molecule drug discovery, 3D shape-based screening, and this study alike, non-CSMs are compounds that medicinal chemists would identify as typical drug-like small molecules of low structural complexity. In contrast, CSMs represent less conventional compounds, characterized, above all, by their larger size (reflected by a high number of heavy atoms and high molecular weight), and along with it, larger numbers of conformational degrees of freedom and/or higher 3D shape complexity (Figure 1). CSMs include, in particular, complex natural products and macrocyclic compounds, many of which are of high relevance to drug discovery but typically lack experimental data. Therefore, if it is found in this study that computational approaches based on 3D shape-based alignment are indeed capable of deriving the likely macromolecular targets of CSMs based on data measured for more conventional small molecules, this could open new avenues to support drug discovery efforts in less densely populated, and hence more innovative, areas of the relevant chemical space.

METHODS

Extraction of High-Quality Data from ChEMBL. The ChEMBL database^{12,28} was processed following a protocol inspired by the work of Bosc et al.²⁹ First, any data records matching the following criteria were extracted from ChEMBL:

- (1) Bioactivity record includes a molecular structure (*canonical_smiles* is not null).
- (2) Reported bioactivity is measured on a single protein or a protein complex (i.e., *confidence_score* 7 or 9).
- (3) *data_validity_comment* is null OR “manually validated”.
- (4) *potential_duplicate* is “0”.
- (5) *activity_comment* is not “inconclusive” OR “unspecified” (capitalization ignored).
- (6) *standard_type* is “Kd” OR “Potency” OR “AC50” OR “IC50” OR “Ki” OR “EC50”.
- (7) NOT (*standard_value* is null AND *pchembl_value* is null AND *activity_comment* is not “active” (capitalization ignored)).
- (8) NOT (*standard_relation* “>”, “≥”, or “>>” AND *standard_value* less than 20 000).

This procedure resulted in a total of 1 452 655 data records. A small number of these data records (2157) had concentrations applied to bioactivity measurements reported in $\mu\text{g}\cdot\text{mL}^{-1}$ as opposed to nM; these values were converted into nM. Next, for each compound–target pair, the median bioactivity value was calculated (because compounds may have assigned more than one bioactivity value for one and the same target). Any compounds with a median activity smaller than or equal to 10 000 nM were defined as active, and all other compounds were discarded. This resulted in a total of 481 194 molecules, corresponding to 786 817 bioactivity records.

Processing of Molecular Structures. The molecular structures extracted from ChEMBL as SMILES were imported into MOE³⁰ (parsing failed for one molecule) and prepared

using MOE’s Wash function. Processing included the removal of the minor components of salts, neutralization, and the addition of hydrogen atoms. Any molecules with a molecular weight in the range of 150 to 1500 Da were kept. The molecules were then labeled “CSM” or “non-CSM” according to the following definition (see Results for motivation and discussion of the thresholds): non-CSMs are compounds with 15 to 30 heavy atoms, whereas CSMs include all compounds with 45 to 55 heavy atoms and all macrocycles with 30 to 55 atoms. Compounds consisting of more than 55 heavy atoms were discarded, as were very small compounds (less than 15 heavy atoms) and CSMs with at least one undefined chiral atom (to ensure that stereochemistry is unambiguously defined for all queries).

Next, conformers were generated with OMEGA,^{31,32} a widely applied, systematic, knowledge-based conformer ensemble generator that makes extensive use of fragment libraries. OMEGA features a “default” or “classic” mode, which handles molecules with rings formed by up to nine atoms, and a macrocycle mode, which handles molecules with larger ring systems. A recent benchmark study of commercial conformer ensemble generators identified OMEGA’s classic algorithm as the best commercial tool with respect to both accuracy and speed.³³ Also OMEGA’s macrocycle mode has been shown to obtain good performance on macrocycles.³⁴

For all non-CSMs (knowledge base compounds), ensembles of a maximum of 400 conformers were calculated with OMEGA (the default value is 200 conformers). OMEGA’s classic mode was employed for all non-CSMs without any rings formed by more than nine atoms (the flipper option, which enumerates the stereochemical configurations of undefined chiral atoms, was enabled). OMEGA’s macrocycle mode was employed to generate conformer ensembles for any molecule with rings formed by more than nine atoms (in accordance with the developer’s specifications).

All CSM queries were represented by the lowest energy conformation generated with OMEGA’s classic or macrocycle modes, applying the same ring size cutoffs as for non-CSMs.

The composition of the data set resulting from this processing workflow is reported in Table 1.

Table 1. Composition of Processed Data Set

		Number of compounds	Number of bioactivity records	Number of targets
Complex small molecules (CSMs)	macrocycles	2780	4618	474 ^a
Complex small molecules (CSMs)	nonmacrocycles	10 870	16 640	1164 ^a
Noncomplex small molecules (non-CSMs)	nonmacrocycles	272 640	460 047	3642

^aCorresponding to a total of 1318 unique targets.

Compilation of a Test Set for Target Prediction. A test set of 28 targets was compiled by following a protocol designed to ensure that the selected proteins are diverse and representative of pharmaceutically relevant protein space. Starting from the sorted list of the 39 proteins with the highest number of CSM records in the processed data set (108–730 CSMs per target), a diverse and representative set

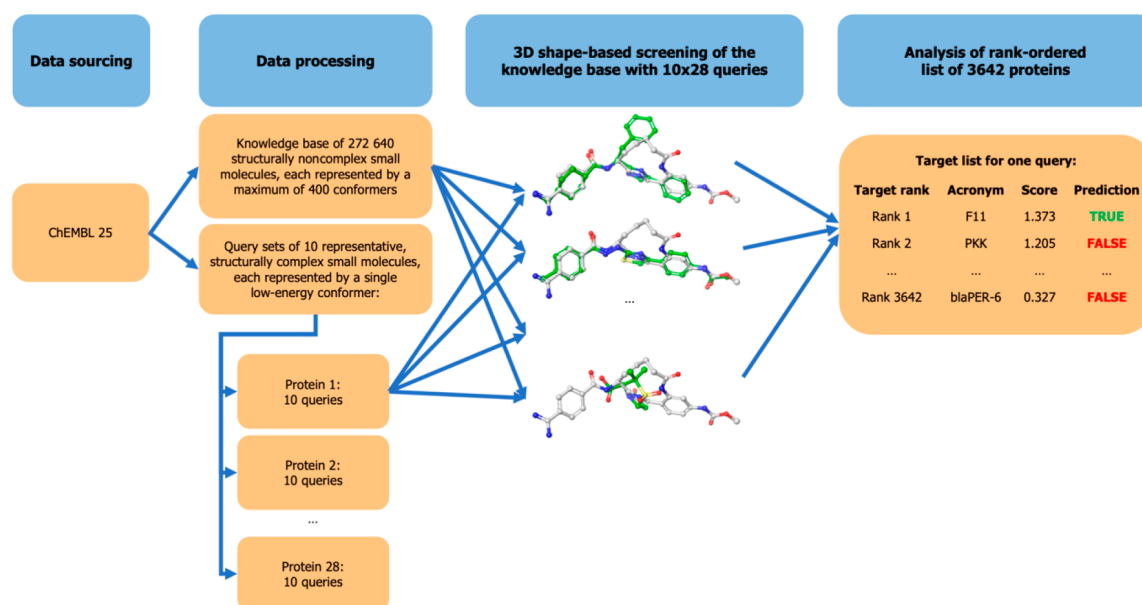


Figure 2. Schematic overview of the general approach.

of proteins was selected based on the following procedure: First, for proteins for which bioactivity records are available for multiple species, only the data for the species with the largest number of CSMs was retained. Second, the protein “protease” from human immunodeficiency virus 1 (CHEMBL2366517) was removed because of the availability of a more comprehensive set of data on the protein “human immunodeficiency virus type 1 protease” (CHEMBL243). Cytochrome P450 enzymes and transporters were excluded because of their wide substrate selectivity and the fact that substrates are known to have multiple binding modes. In the final step, the remaining proteins were clustered with CD-HIT^{35,36} based on their full-length amino acid sequence (a sequence identity cutoff of 0.4 was employed for this procedure). For each of the clusters, only the protein with the largest number of CSMs was kept. With the 28 targets of interest now defined, in the next step, for each of the selected proteins, the 10 most diverse CSMs were determined with MOE’s function for the generation of diverse subsets (using MACCS fingerprints in combination with the Tanimoto coefficient).

Target Prediction. The 280 (28 × 10) CSMs served as queries for screening with ROCS^{37,38} against the knowledge base of 272 640 non-CSMs (note that the number of unique CSMs is 269 as a minority of the selected CSMs are active on more than one of the selected 28 proteins). The proteins were ranked according to the maximum similarity between a CSM query and all non-CSM ligands recorded for a protein in the knowledge base.

Molecular similarity was quantified separately by each of four similarity metrics implemented in ROCS: ShapeTanimoto, TanimotoCombo, RefTverskyCombo, and FitTverskyCombo score. As suggested by their names, metrics are either based on the Tanimoto or the Tversky coefficient. The Tanimoto coefficient quantifies the similarity of two molecules, f and g , based on their self-volume overlaps (I_f and I_g) and the volume overlap between the two molecules ($O_{f,g}$)

$$\text{Tanimoto}_{f,g} = \frac{O_{f,g}}{I_f + I_g - O_{f,g}}$$

The Tversky coefficient can be asymmetric (depending on the α and β parameters chosen), hence allowing to emphasize on either substructure or superstructure matching

$$\text{Tversky}_{f,g} = \frac{O_{f,g}}{\alpha I_f + \beta I_g}$$

The ShapeTanimoto score ranges from 0 to 1, with a value of 1 indicating a perfect fit of molecular shapes. Importantly, the ShapeTanimoto score only considers the fit of shapes for the volume overlap, whereas the three “combo” scores additionally take the type and distribution of chemical features into account. The “combo” scores typically range from 0 to 2, with equal weights applied to the shape and color components.

The RefTverskyCombo score assigns an α value of 0.95 to the CSM query molecule as the main self-overlap term, meaning, in the context of this study, that it emphasizes the matching of the CSM (which, by design of the data sets, is the superstructure). The FitTverskyCombo score, on the contrary, assigns a β value of 0.95 to the fit molecule (i.e., the knowledge base molecule), emphasizing the match of the non-CSM (substructure). Note that the RefTverskyCombo and FitTverskyCombo scores can have values greater than 2 because the overlap of two compounds can be larger than a molecule’s self-overlap.

ROCS was run with factory settings with the following exceptions: both “-besthits” and “-maxhits” were set to “0” in order to cause ROCS to retain all results. The “-rankby” option was set to an appropriate value in order to have the results ranked by the four similarity metrics. For experiments using the ShapeTanimoto score, the “-shapeonly” function was enabled in order to cause ROCS to align molecules by taking only molecular shape into account (and not color). Targets assigned identical scores were also assigned identical ranks.

Table 2. Overview of Targets Selected for Testing Performance of 3D Shape-Focused Target Prediction Approach

Target ID	Target name	Protein classification	Target abbreviation	Organism	No. CSMs ^a	No. non-CSMs ^b
CHEMBL243	Human immunodeficiency virus type 1 protease	enzyme	HIV-1 protease	Human immunodeficiency virus 1	703	185
CHEMBL2362980	Paired box protein Pax-8	unclassified	PAX8	<i>Homo sapiens</i>	390	465
CHEMBL270	Mu opioid receptor	membrane receptor	MOR	<i>Rattus norvegicus</i>	337	299
CHEMBL4616	Ghrelin receptor	membrane receptor	GHSR	<i>Homo sapiens</i>	299	127
CHEMBL2001	Purinergic receptor P2Y12	membrane receptor	P2Y12	<i>Homo sapiens</i>	290	70
CHEMBL4822	Beta-secretase 1	enzyme	BACE1	<i>Homo sapiens</i>	289	1634
CHEMBL3717	Hepatocyte growth factor receptor	enzyme	HGFR	<i>Homo sapiens</i>	274	800
CHEMBL3948	Angiotensin II type 1a (AT-1a) receptor	membrane receptor	AGTR1	<i>Oryctolagus cuniculus</i>	266	43
CHEMBL4860	Apoptosis regulator Bcl-2	ion channel	BCL2	<i>Homo sapiens</i>	266	84
CHEMBL203	Epidermal growth factor receptor erbB1	enzyme	EGFR	<i>Homo sapiens</i>	233	1451
CHEMBL259	Melanocortin receptor 4	membrane receptor	MC4R	<i>Homo sapiens</i>	233	85
CHEMBL325	Histone deacetylase 1	epigenetic regulator	HDAC1	<i>Homo sapiens</i>	192	1453
CHEMBL1957	Insulin-like growth factor I receptor	enzyme	IGF1R	<i>Homo sapiens</i>	177	514
CHEMBL2820	Coagulation factor XI	enzyme	F11	<i>Homo sapiens</i>	173	15
CHEMBL5023	p53-binding protein Mdm-2	other nuclear protein	MDM2	<i>Homo sapiens</i>	156	183
CHEMBL5658	Prostaglandin E synthase	enzyme	PGES	<i>Homo sapiens</i>	153	288
CHEMBL5251	Tyrosine-protein kinase BTK	enzyme	BTK	<i>Homo sapiens</i>	147	83
CHEMBL286	Renin	enzyme	REN	<i>Homo sapiens</i>	144	84
CHEMBL4414	Plasmeprin 2	enzyme	PM2	<i>Plasmodium falciparum</i>	144	15
CHEMBL220	Acetylcholinesterase	enzyme	AChE	<i>Homo sapiens</i>	130	1083
CHEMBL2327	Neurokinin 2 receptor	membrane receptor	NK2R	<i>Homo sapiens</i>	129	45
CHEMBL2954	Cathepsin S	enzyme	CTSS	<i>Homo sapiens</i>	123	424
CHEMBL4662	Proteasome Macropain subunit MB1	enzyme	MB1	<i>Homo sapiens</i>	121	73
CHEMBL240	HERG	ion channel	HERG	<i>Homo sapiens</i>	117	2260
CHEMBL244	Coagulation factor X	enzyme	F10	<i>Homo sapiens</i>	115	277
CHEMBL3572	Cholesteryl ester transfer protein	ion channel	CETP	<i>Homo sapiens</i>	114	26
CHEMBL1865	Histone deacetylase 6	epigenetic regulator	HDAC6	<i>Homo sapiens</i>	112	1070
CHEMBL3706	ADAM17	enzyme	ADAM17	<i>Homo sapiens</i>	108	256

^aNumber of ligands that are CSMs. ^bNumber of ligands that are non-CSMs.

RESULTS AND DISCUSSION

The aim of this work is to determine the capacity of 3D alignment-dependent shape-based approaches to predict the macromolecular targets of CSMs based on their similarity to non-CSMs with measured bioactivities (Figure 2).

Defining what constitutes a complex or a noncomplex molecule is a nontrivial task because molecular complexity is context dependent and its perception inherently subjective. Thus, it does not come as a surprise that there is no universally applicable and easily interpretable metric for the quantification of molecular complexity in existence.³⁹

Our aim was to identify an effective, robust, and, importantly, easily interpretable metric. We investigated several of the many complexity metrics discussed in a recent review.³⁹ By visual inspection of the molecular structures contained in our processed data sets, we unanimously converged on using the number of heavy atoms as a metric of structural complexity for the following reasons:

- (1) The number of heavy atoms correlates well with molecular weight (and molecular size), the most

important parameter in drug discovery besides log *P*, and chemists are well familiar with it.

- (2) In the context of shape-based screening, the number of heavy atoms is more descriptive of molecular complexity than other common measures such as the number (or fraction) of Csp³ atoms because nonplanarity itself does not pose a particular challenge to the algorithms under investigation.
- (3) The aim of this study is to understand the limits of 3D shape-based approaches for target prediction, and these are, like for most other in silico approaches, defined primarily by the available data, and there are clearly more data available for conventional drug-like compounds (small “small molecules” with molecular weight below 500 Da), than there are for larger-sized compounds (Figure S1).

Hence, for the purpose of this study, non-CSMs are any compounds consisting of 15–30 heavy atoms (corresponding to an average molecular weight from 222 to 424 Da for this data set). In contrast, CSMs are compounds that are unusually large (minimum of 45 heavy atoms; corresponding to an average of 631 Da) or macrocyclic with at least 30 heavy

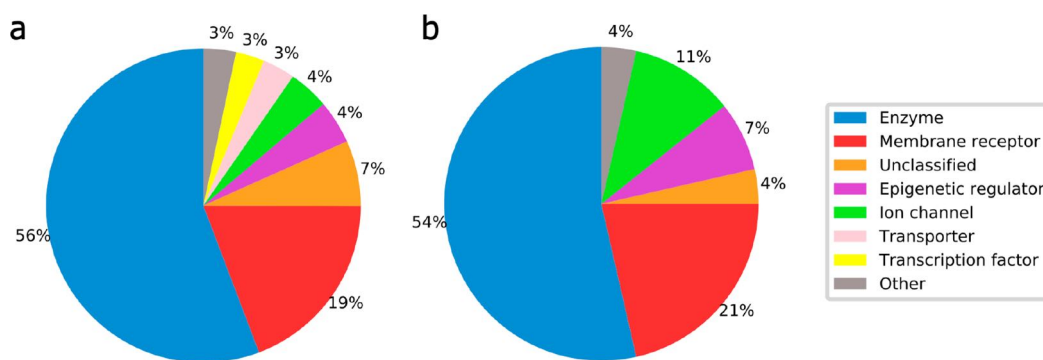


Figure 3. Comparison of the distribution of target classes across (a) all (1318) proteins with at least one known CSM ligand and (b) the 28 targets selected for this study.

atoms. Any compounds with more than 55 heavy atoms (corresponding to an average molecular weight of 772 Da) were not considered in this study because of the excessive size of their conformational space. The numbers of CSMs and non-CSMs present in the processed ChEMBL data set are reported in Table 1.

Twenty-eight representative and pharmaceutically relevant targets were selected for testing, each represented by the 10 most diverse bioactive CSMs (giving rise to a total of 280 CSM queries). Each of the 280 CSM queries was represented by a calculated minimum energy conformation, whereas each of the 272 640 non-CSMs of the knowledge base (with measured bioactivities on a total of 3642 proteins) was represented by up to 400 conformers representative of the low-energy conformational space.

Characterization of Data Sets Underlying the Evaluation. Targets. The 28 targets selected for this study (Table 2) are diverse and a good representation of the pharmaceutically relevant protein space. The pairwise identity of the full-length protein sequence of all selected targets is below 40%. Most target classes are well represented, as shown by the comparison of the target class distributions over all proteins that have at least one CSM ligand (1318 proteins) and the 28 selected targets (Figure 3). Only transporters and transcription factors are not represented. The transporters represented by a significant number of diverse CSMs in the data set bind a wide variety of substrates, in part with clearly distinct binding modes, for which reason we excluded them, as we excluded cytochrome P450 3A4 for the same reason. There are no transcription factors with sufficient numbers of CSM records that would allow their inclusion in this study.

Complex and Noncomplex Small Molecules. The physicochemical property spaces of the 13 650 CSMs and 272 640 non-CSMs serving as the data basis of this work are clearly distinct, as shown in Figure 4. While most CSMs in this study have a molecular weight between 550 and 800 Da (median 664 Da), most non-CSMs have a molecular weight of less than 500 Da (median 355 Da; Figure 4a). Analogous observations are made for the number of heavy atoms (Figure 4b), where the median is 47 for CSMs and 25 for non-CSMs. CSMs have a substantially higher number of rotatable bonds than non-CSMs (median 11 vs 4; Figure 4c) and also a higher number of chiral centers on average (median 2 vs 0; Figure 4d). Also the average number of rings (Figure 4e) and the number of aromatic rings (Figure 4f) are higher for CSMs (average 4.96 and 3.39, respectively) than for non-CSMs (average 3.23 and 2.46, respectively). Although the fraction of

heteroatoms (Figure 4g) in CSMs and non-CSMs is comparable (median 0.25 for both classes of compounds), the log P (Figure 4h) is higher for CSMs (median 4.85 and 3.33, respectively).

Performance of Shape-Based Screening with Different Similarity Metrics. ROCS features two different alignment modes: a default mode, which takes into account both molecular shape and color, and the shape-only mode, which considers molecular shape only. Both of these alignment modes were assessed in this study with different scores implemented in ROCS in the following setups (consistent with the underlying algorithm): (i) the default alignment mode in combination with the TanimotoCombo, RefTverskyCombo, and FitTverskyCombo scores and (ii) the ShapeTanimoto score in combination with ROCS' shape-only mode (i.e., with the -shapeonly function enabled).

Performance Measured for Individual Complex Small Molecules. Among the four investigated scores, the TanimotoCombo score clearly outperformed all other scores in ranking the targets of CSMs among the top positions of 3642 proteins (Table 3 and Figure 5a; note for the figure that steeper curves indicate worse performance and that the y -axis is on a logarithmic scale). With the TanimotoCombo score, the target of interest (i.e., the target assigned to this particular query) was ranked among the top-5 positions for 83 (30%) of the 280 CSM queries (note that the automated query selection procedure resulted in the selection of 10 CSMs which are active on more than one of the 28 targets; accordingly, these CSMs represent more than one query). The success rate increases to 41% when considering the top-20 ranks and to 47% when considering the 40 top-ranked proteins (which corresponds to roughly 1% of the total list of proteins represented by the knowledge base).

Compared to the TanimotoCombo score, the success rates obtained by the ShapeTanimoto, RefTverskyCombo, and FitTverskyCombo scores were roughly 20 percentage points lower. The RefTverskyCombo score tended to have higher success rates than the ShapeTanimoto and FitTverskyCombo scores when considering a greater number of ranks (top-40, top-80, and top-200).

In order to obtain a better understanding of the reasons for the observed differences in the target ranking performance of the individual scores, we (i) visually inspected alignments and related them to the respective score values, (ii) analyzed the relationships between scores and ranks, and (iii) determined the relationships between scores and molecular weight.

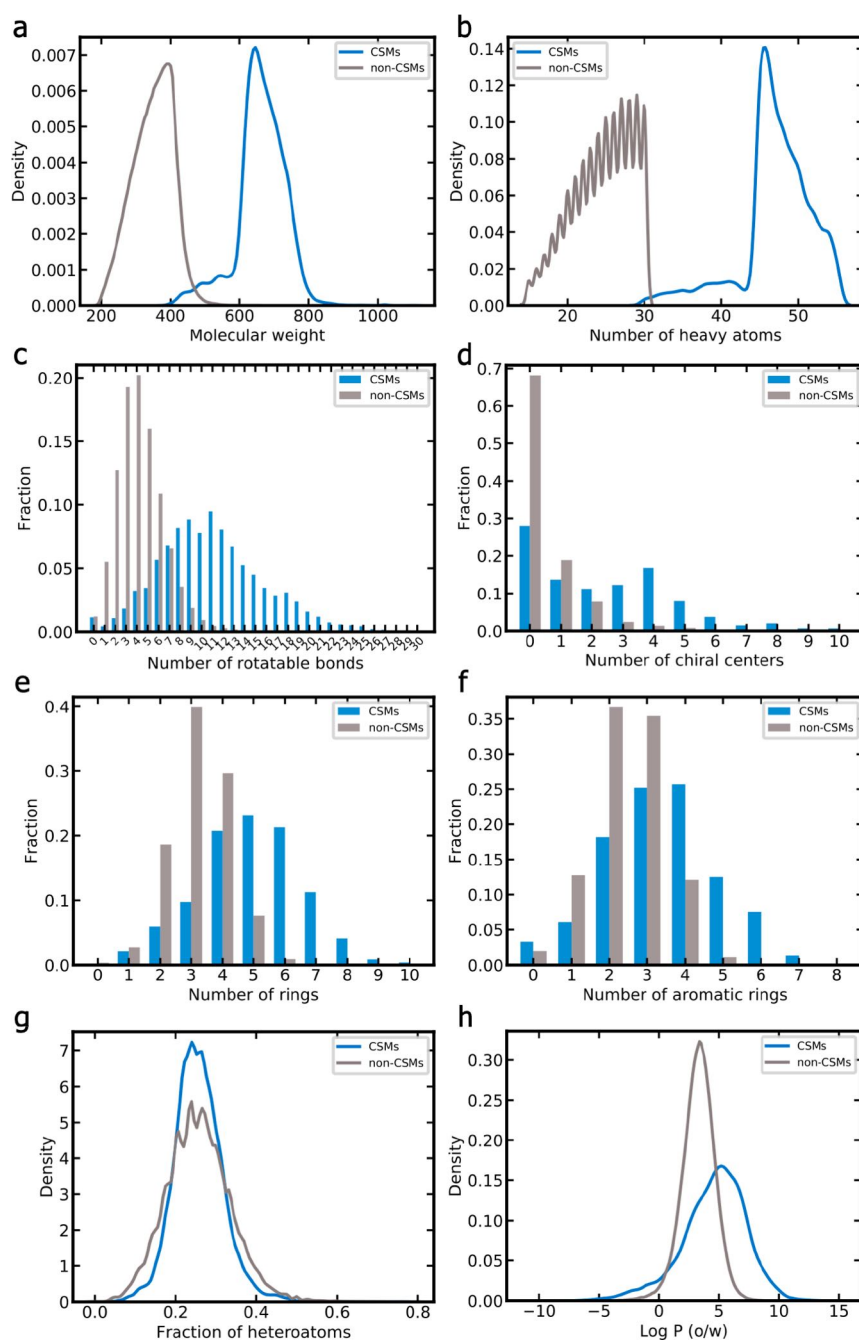


Figure 4. Comparison of the physicochemical property spaces of CSMs (blue) and non-CSMs (gray): (a) molecular weight, (b) number of heavy atoms, (c) number of rotatable bonds, (d) number of chiral centers, (e) number of rings, (f) number of aromatic rings, (g) fraction of heteroatoms, and (h) $\log P$.

Table 3. Success Rates for Predicting Targets of Interest of Queries with Different Scoring Functions

Rank	All/macrocylic/nonmacrocylic complex small molecules (CSMs) [%]			
	TanimotoCombo score	ShapeTanimoto score	RefTverskyCombo score	FitTverskyCombo score
Top-5	30/20/31	9/2/10	11/7/12	9/4/10
Top-10	37/27/39	14/7/16	12/9/12	11/4/12
Top-20	41/29/43	20/11/22	22/13/23	14/7/15
Top-40 (~1%)	47/33/49	24/11/27	35/18/38	19/7/22
Top-80	54/42/56	34/20/37	46/24/51	28/16/30
Top-200	62/60/63	51/36/54	60/58/60	46/42/47

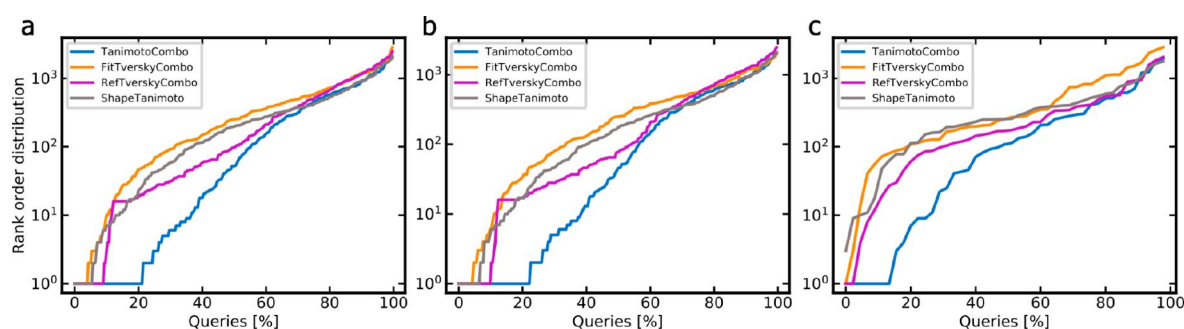


Figure 5. Percentage of queries for which the target of interest (out of 3642 proteins) was assigned ranks better than or equal to the ranks indicated on the *y*-axis (“rank order distribution”) for (a) all queries, (b) nonmacrocyclic queries, and (c) macrocyclic queries. Note that steeper curves indicate worse performance and that the *y*-axis is on a logarithmic scale.

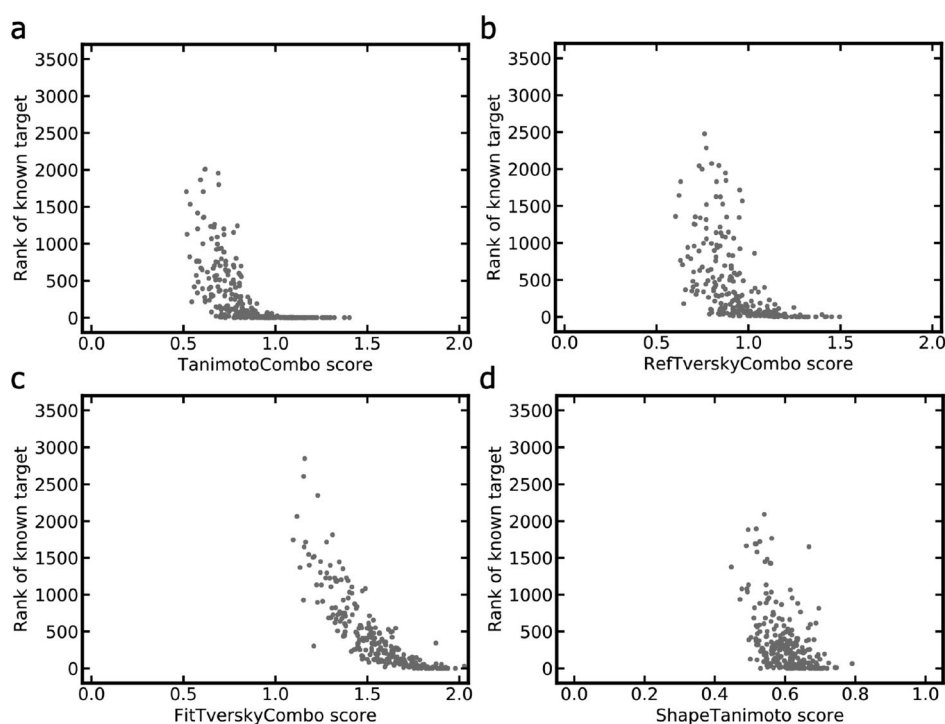


Figure 6. Relationship between the (a) TanimotoCombo, (b) RefTverskyCombo, (c) FitTverskyCombo, and (d) ShapeTanimoto scores and the ranks obtained for the targets of interest of the 280 CSM queries. Note that there is one instance where the FitTverskyCombo score is greater than 2.0 (see [Target Prediction](#) section in the [Methods](#) section for an explanation).

The FitTverskyCombo score emphasizes the matching of the knowledge base molecule (which is the smaller-sized molecule in this context). We found that the parametrization of the FitTverskyCombo score leads to the preference for knowledge base molecules that are particularly small in size because there is a high likelihood for these molecules to produce good matches with a part of the CSM. This preference is reflected by negative Pearson’s and Spearman’s correlation coefficients for the FitTverskyCombo score and molecular weight (−0.37 and −0.39, respectively; numbers report averages over all CSM queries). The fact that alignments of CSMs with small non-CSMs have a high likelihood of obtaining high FitTverskyCombo scores is visible from [Figure 6](#), where it is shown that the FitTverskyCombo function indeed assigns high scores to a much larger proportion of CSMs aligned with their nearest non-CSM ([Figure 6c](#)) than any of the other scoring functions ([Figure 6a, b, d](#)). This

behavior results in high false-positive prediction rates of this score in the study context, which explains the inferior performance over the TanimotoCombo score.

The RefTverskyCombo score emphasizes the matching of the CSM and, consequently, has a preference for larger molecules, which is reflected by averaged Pearson’s and Spearman’s correlation coefficients of 0.43 and 0.40, respectively. Consistent with the fact that pairs of larger-sized molecules are less likely to produce good matches, the proportion of targets for which the best match is assigned a high RefTverskyCombo score value is substantially lower than for the FitTverskyCombo score ([Figure 6b, c](#)).

The reason for the superior performance of the TanimotoCombo score appears to be the fact that, as a balanced measure of molecular similarity, its ranking capacity is less affected by differences in the size of molecules. This is reflected by lower averaged Pearson’s and Spearman’s correlation coefficients

between the score and molecular weight (0.39 and 0.33, respectively). Figure 6a shows that high TanimotoCombo scores generally go along with high target ranks (observed as a tail toward the bottom right corner of the plot), which is often not the case for other scores, in particular, the FitTversky-Combo and ShapeTanimoto scores.

The obvious explanation for the inferior performance of the ShapeTanimoto score over the three “combo” scores is the neglect of chemistry, which leads to a lack of specificity during alignment and scoring and, in turn, a clear preference for matches involving larger-sized non-CSMs (averaged Pearson’s and Spearman’s correlation coefficients 0.62 and 0.51, respectively). ShapeTanimoto scores are often high (Figure 7) because good overlaps of molecular shapes are likely when

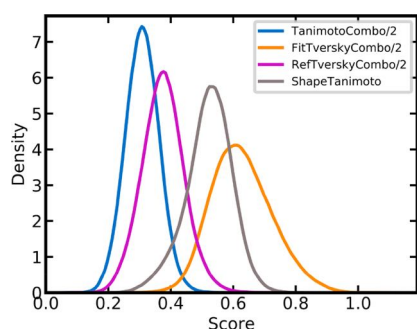


Figure 7. Density distributions of the four similarity metrics over all lists of scores obtained for all 280 queries. The TanimotoCombo, RefTverskyCombo, and FitTverskyCombo score values were scaled to the same range as the ShapeTanimoto score.

chemical features (color) are not considered. However, high ShapeTanimoto scores often do not correspond to high target rankings (Figure 6d), which is another indication of the lack of specificity of this score.

Further conclusions that can be derived from these analyses are that values obtained with different scores should not be directly compared. Moreover, the scores obtained for individual query–target combinations should not be used as a measure of the likelihood of a compound to be active on that target. In other words, the predictions provide an indication of the likelihood of a protein being a target only relative to all other possible targets.

Performance Measured on a Per-Target Basis. A further way of analyzing success rates is on a per-target basis, evaluating the results for query sets (the 10 queries) rather than individual queries. For 24 of the 28 targets (86%), the TanimotoCombo score assigned the top rank to the target of interest for at least one of the 10 queries (Figure 8). For the ShapeTanimoto, RefTverskyCombo, and FitTverskyCombo scores, this was only the case for 43%, 57%, and 29% of the 28 proteins, respectively. Additional details are provided in Table 4.

Only for four out of 28 targets, the TanimotoCombo score failed to rank the target of interest among the top-10 positions with any of the 10 queries: the paired box protein Pax-8 (*Homo sapiens*), plasmepsin 2 (*Plasmodium falciparum*), neurokinin 2 receptor (*Homo sapiens*), and cholesteryl ester transfer protein (*Homo sapiens*).

For the paired box protein Pax-8, the highest rank obtained with any of the 10 queries was 32 (TanimotoCombo score). One of the reasons for failure is the fact that most of the CSMs

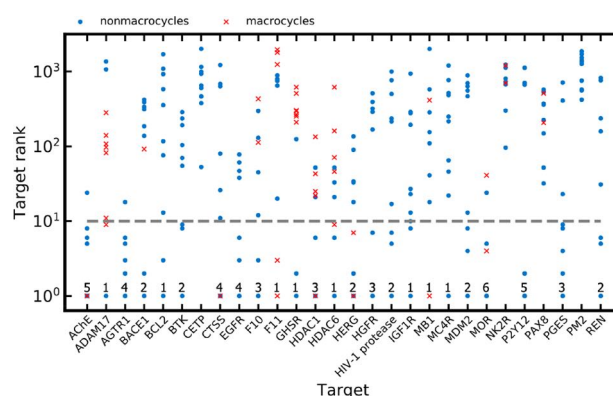


Figure 8. Ranks assigned with the TanimotoCombo score to the target of interest for the 280 CSM queries. Note that the y-axis is on a logarithmic scale. The numbers reported at the bottom of the graph indicate the number of CSM queries for which the target of interest was assigned the rank of 1 (indicating perfect prediction); the dashed line indicates the rank of 10.

active on this target are very different from the bioactive non-CSMs in terms of chemistry. They are characterized by long and flexible scaffolds; a minority are macrocyclic (indicated in Figure 8).

In the case of plasmepsin 2, the best rank obtained was just 420 (TanimotoCombo score). This target is characterized by a highly flexible ligand binding site to which small molecules are known to bind in several distinct modes.⁴⁰ The fact that there were only 15 non-CSMs recorded for that target may contribute to the difficulties in recognizing CSMs active on this protein (note, however, that coagulation factor XI was correctly identified as the target of two out of the 10 CSMs and ranked among the top-3 positions even though the target is represented by only 15 non-CSMs in the knowledge base).

For the neurokinin 2 receptor, the best rank obtained with any of the 10 CSMs was 96 (TanimotoCombo score). The reasons for failure appear to be similar to those for Pax-8. Most of the CSMs have a substantial number of rotatable bonds; a minority are macrocyclic.

For the cholesteryl ester transfer protein, the best rank obtained with any of the 10 CSMs was 53 (TanimotoCombo score). The CSM queries of the cholesteryl ester transfer protein are characterized by three to four similarly sized branches originating from a central carbon or nitrogen atom. The structures of most CSM queries are clearly distinct from those of the ligands represented in the knowledge base.

Overall, the results obtained on a per-target basis indicate that the value of the method can be substantially higher in cases where several compounds targeting the same protein are explored, although this scenario is rare in the context of CSMs (as opposed to conventional drug-like compounds). A further conclusion (derived from the results presented in Figure 8) is that there is no correlation between the success rates for a target and the number of non-CSM representing that target in the knowledge base.

Performance on Macrocyclic as Compared to Non-macrocyclic Complex Small Molecules. Forty-five of the 280 CSMs are macrocyclic, covering 14 out of the 28 targets studied in this work. The ring systems of the 45 macrocyclic CSMs are formed by up to 22 atoms, with a median of 15 atoms (Figure 9).

Table 4. Best and Median Target Ranks Obtained by Different Scores for Query Sets Consisting of 10 CSMs Each

Protein ^a	Target rank with score							
	TanimotoCombo		RefTverskyCombo		FitTverskyCombo		ShapeTanimoto	
	best	median	best	median	best	median	best	median
HIV-1 protease	1.0	116.0	1.0	135.0	2.0	381.5	7.0	356.0
PAX8	32.0	294.0	83.0	315.0	80.0	216.0	126.0	253.0
MOR	1.0	1.0	16.0	19.5	12.0	88.0	1.0	34.0
GHSR	1.0	260.0	1.0	213.5	11.0	794.0	4.0	349.0
P2Y12	1.0	1.5	1.0	24.0	1.0	67.0	1.0	185.5
BACE1	1.0	162.0	16.0	320.0	32.0	304.5	54.0	197.0
HGFR	1.0	87.5	1.0	84.5	6.0	162.5	1.0	59.0
AGTR1	1.0	2.0	1.0	2.0	3.0	89.5	2.0	20.5
BCL2	1.0	236.5	16.0	188.5	153.0	705.0	1.0	280.5
EGFR	1.0	4.5	1.0	18.0	1.0	69.5	1.0	59.0
MC4R	1.0	233.0	28.0	475.5	25.0	274.0	1.0	289.5
HDAC1	1.0	21.5	1.0	63.0	1.0	96.0	1.0	78.5
IGF1R	1.0	25.0	1.0	29.0	1.0	310.0	1.0	126.5
F11	1.0	774.0	1.0	901.0	139.0	1765.0	1.0	462.5
MDM2	1.0	240.5	2.0	326.0	3.0	235.0	1.0	143.5
PGES	1.0	6.0	1.0	41.0	3.0	285.5	8.0	96.0
BTK	1.0	62.5	1.0	59.0	1.0	652.0	1.0	200.0
REN	1.0	95.0	1.0	187.0	1.0	673.5	161.0	599.0
PM2	420.0	1308.5	534.0	1257.0	636.0	1225.0	440.0	1452.0
AChE	1.0	3.0	1.0	47.5	1.0	29.5	17.0	41.0
NK2R	96.0	712.0	305.0	908.5	83.0	372.5	287.0	921.5
CTSS	1.0	18.5	1.0	64.0	1.0	88.0	4.0	99.0
MB1	1.0	132.5	8.0	116.5	17.0	136.0	5.0	529.5
HERG	1.0	12.5	1.0	49.0	28.0	81.5	13.0	62.0
F10	1.0	28.5	16.0	74.5	10.0	420.5	1.0	58.5
CETP	53.0	625.0	1063.0	1772.0	93.0	443.5	6.0	484.0
HDAC6	1.0	39.5	16.0	84.5.0	5.0	89.5	11.0	166.0
ADAM17	1.0	102.5	1.0	141.0	4.0	229.0	2.0	222.0

^aFor the explanation of all target acronyms, see Table 2.

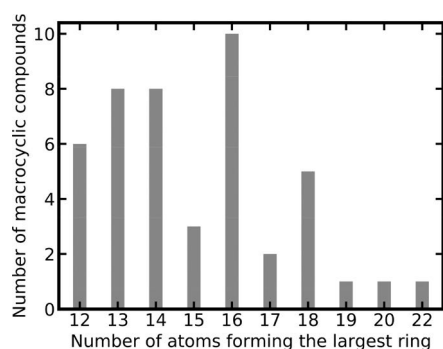


Figure 9. Size of largest ring systems of 45 macrocyclic CSMs.

Our results show that the task of target prediction is more challenging for macrocyclic compounds than for nonmacrocyclic ones (Figure 5b, c). For the TanimotoCombo score, the top-5, top-10, top-20, and top-40 success rates for non-macrocyclic CSMs were 31%, 39%, 43%, and 49%, respectively, whereas for macrocyclic CSMs, they were just 20%, 27%, 29%, and 33%, respectively. Besides the low molecular similarity of macrocyclic compounds with the non-CSMs of the knowledge base, a major reason for the lower success rates observed for macrocyclic compounds are the complexities involved in representing the 3D conformations of these queries, related to a high number of conformational degrees of freedom and

torsional properties that are distinct from nonmacrocyclic compounds.

Cases Where at Least One Score Worked Well While Others Failed. There are several examples of CSMs for which their targets were ranked at high positions with one score while other scores failed. We identified nine CSMs (three of them being macrocyclic compounds) for which their targets were assigned ranks of 10 or better by at least one score while other score(s) assigned ranks of 450 or worse (Table 5). In seven out of the nine cases, the TanimotoCombo score performed well, while others failed (Figure 10a, b); in two cases the ShapeTanimoto score outperformed the other scores (Figure 10c, d). For the examples reported in Table 5, it can be seen that the alignments produced by the three “combo” scores are generally more consistent in terms of chemistry (in particular, with regard to the orientation of chemical features) than the alignments produced by the ShapeTanimoto score. However, the FitTverskyCombo score failed to identify the target of interest for many CSMs due to its emphasis on matching the knowledge base molecule (substructure; see Performance of Shape-Based Screening with Different Similarity Metrics section in the Results section). In contrast, the ShapeTanimoto score often failed because of its disregard of chemistry, which is reflected by alignments that lack the matching of chemical features.

Performance as a Function of Molecular Similarity. The performance of similarity-based approaches depends on

Table 5. Examples of CSMs for Which Their Targets Were Successfully Identified by One at Least One Score While Others Failed

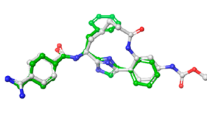
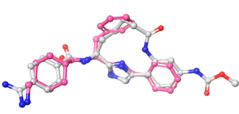
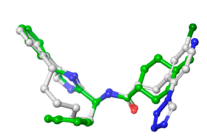
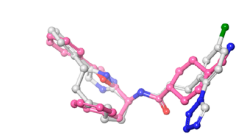
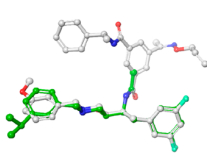
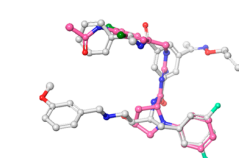
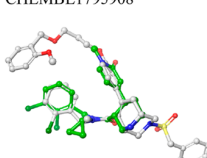
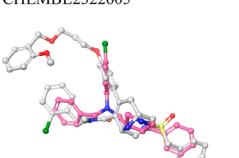
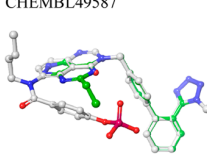
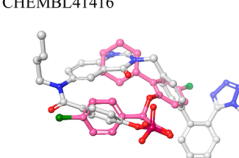
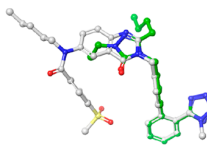
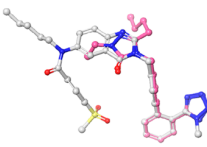
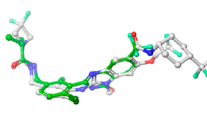
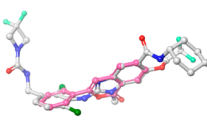
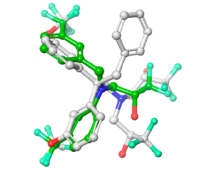
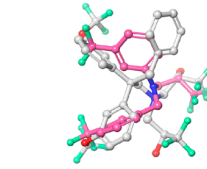
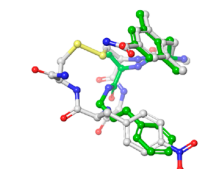
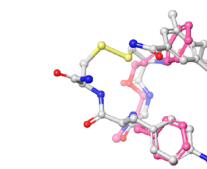
Query ^a	Target ^b	Rank by score				Alignments of CSM queries and reference compounds obtaining the	
		Tanimoto Combo	FitTversky Combo	RefTversky Combo	Shape Tanimoto	highest rank ^c	lowest rank ^c
CHEMBL 3699200*	F11	1	208	1	1649	TanimotoCombo score CHEMBL3393362; CHEMBL3355664	ShapeTanimoto score CHEMBL3355686
							
CHEMBL 3676156*	F11	3	549	27	815	TanimotoCombo score CHEMBL3393362; CHEMBL3355664	ShapeTanimoto score CHEMBL3355685
							
CHEMBL 553424	BACE1	2	32	23	578	TanimotoCombo score CHEMBL1760861	ShapeTanimoto score CHEMBL3627959
							
CHEMBL 508748	REN	5	180	79	561	TanimotoCombo score CHEMBL1795908	ShapeTanimoto score CHEMBL2322605
							
CHEMBL 281890	AGTR1	5	551	17	69	TanimotoCombo score CHEMBL49587	FitTverskyCombo score CHEMBL41416
							

Table 5. continued

Query ^a	Target ^b	Rank by score				Alignments of CSM queries and reference compounds obtaining the	
		Tanimoto Combo	FitTversky Combo	RefTversky Combo	Shape Tanimoto	highest rank ^c	lowest rank ^c
CHEMBL 27903	AGTR1	3	535	16	205	TanimotoCombo score CHEMBL86084	FitTverskyCombo score CHEMBL86084
							
CHEMBL 3694569	PGES	9	471	65	92	TanimotoCombo score CHEMBL3342705	FitTverskyCombo score CHEMBL2140153
							
CHEMBL 3683924	CETP	53	93	1351	6	ShapeTanimoto score CHEMBL340397	RefTverskyCombo score CHEMBL340397
							
CHEMBL 445869*	MOR	41	461	85	3	ShapeTanimoto score CHEMBL171763	FitTverskyCombo score CHEMBL2048969
							

^aQueries marked with a "*" are macrocyclic compounds. ^bF11, coagulation factor XI; BACE1, beta-secretase 1; REN, renin; AGTR1, angiotensin II type 1a (AT-1a) receptor; PGES, prostaglandin E synthase; CETP, cholesteryl ester transfer protein; MOR, mu opioid receptor. ^cChEMBL IDs reported are those that obtained the highest/lowest rank for the target of interest of the individual CSM queries, according to the scoring function indicated in the respective table cells. Alignments shown are those that obtained the highest rank for a CSM query. In cases where multiple alignments obtained identical scores (and ranks), only one alignment is shown.

how well the query is represented by the data stored in the knowledge base. In the context of this study, one of the simplest measures of the molecular similarity is the difference in the number of heavy atoms between the CSM query and the nearest non-CSM ligand. Figure 11a and b shows that the success rates of the method are largely unaffected by the differences in the number of heavy atoms over the observed range. The compatibility of chemical features seems to play a much more important role than pure differences in molecular size. This is confirmed when using the Tanimoto coefficient derived from 2D Morgan2 fingerprints as a measure of molecular similarity. As shown in Figure 11c, ROCS (in

combination with the TanimotoCombo score) ranked 43% of all CSMs with a maximum Tanimoto coefficient between 0.2 and 0.3 among the top-10 positions and 73% of all CSMs with a coefficient between 0.3 and 0.4. This robustness is remarkable, as molecular structures with a Morgan2 fingerprint-based Tanimoto coefficient below 0.4 are clearly distinct in most cases. Importantly, it is likely that compounds with such a low degree of molecular similarity have different binding modes, which is beyond the reach of any ligand-based approach.

Among the 280 queries investigated in this work, we identified 11 compounds (six of them are macrocyclic

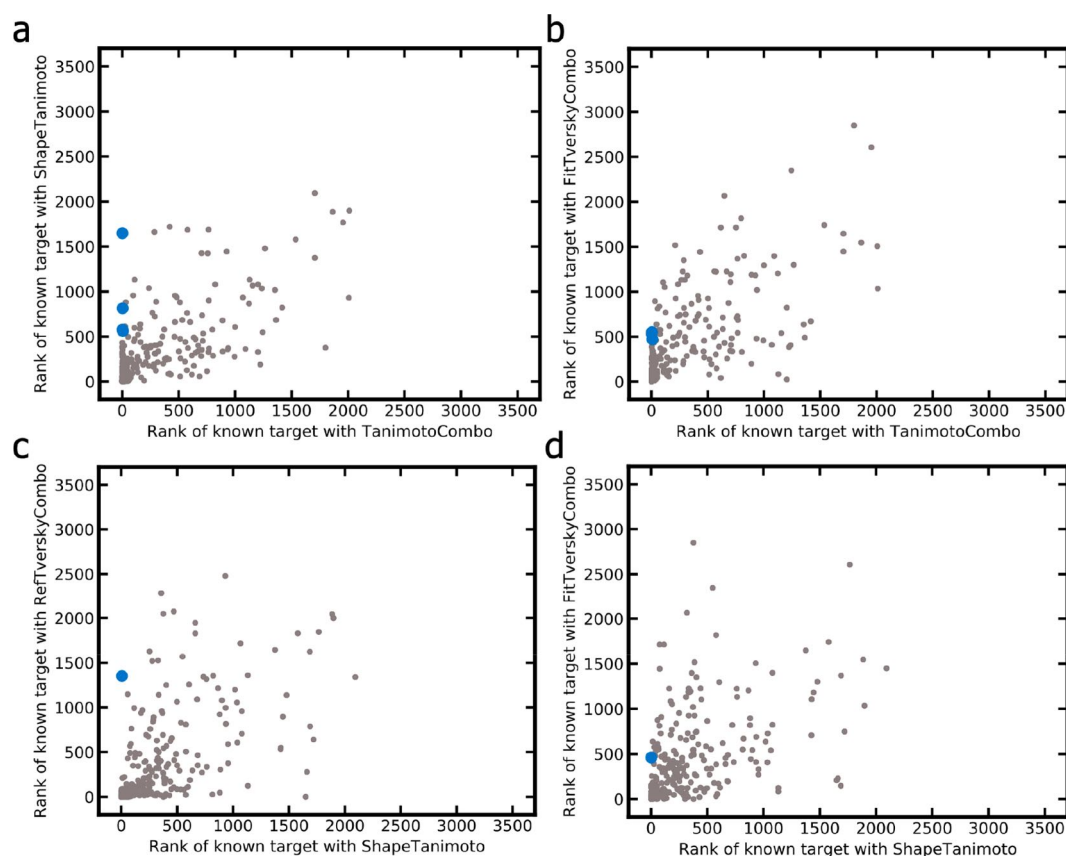


Figure 10. Ranks assigned to the targets of interest of the 280 CSM queries by the (a) TanimotoCombo vs ShapeTanimoto scores, (b) TanimotoCombo vs FitTverskyCombo scores, (c) ShapeTanimoto vs RefTverskyCombo scores, and (d) ShapeTanimoto vs FitTverskyCombo scores. The nine compounds for which one score produced good results while others failed are highlighted in blue.

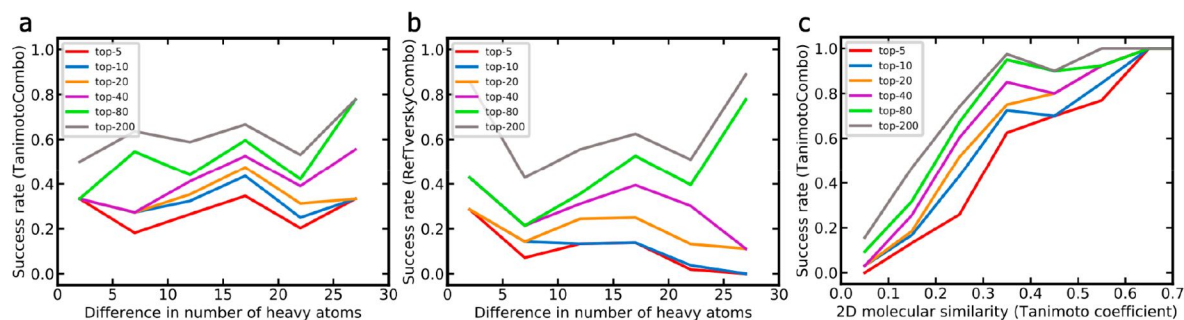


Figure 11. Success rates (i.e., fraction of CSM queries for which the target of interest was ranked among the top-*k* positions) and how they are influenced by the structural relationship between the query CSM and the nearest ligand (non-CSM) recorded in the knowledge base: (a) success rates of the TanimotoCombo score as a function of the difference of molecular size (quantified as number of heavy atoms, separated into bins of size 5), (b) success rates of the RefTverskyCombo score as a function of the difference of molecular size (separated into bins of size 5), and (c) success rates of the TanimotoCombo score as a function of the 2D molecular similarity quantified as Tanimoto coefficient based on Morgan2 fingerprints (separated into bins of size 0.1). Note that in panel (c) success rates for queries with a Tanimoto coefficient greater than 0.7 are not reported because of the limited number of examples. The trends observed in panel (c) are consistent with those observed when using atom type fingerprints instead of Morgan2 fingerprints to quantify 2D molecular similarity and also when using the Tversky coefficient ($\alpha = 0.95$) instead of the Tanimoto coefficient (data not shown).

compounds) for which their target was ranked within the top-10 positions out of 3642 targets, despite being structurally extremely dissimilar from any ligands (non-CSMs) recorded in the knowledge base (Tanimoto coefficients lower than 0.18). As shown in Table 6, most of the alignments produced by ROCS for the 11 compounds are not only plausible and sensible from a chemistry point of view but also visually easily

interpretable thanks to the hard Gaussians used by ROCS for chemical features (color), which cause a lock-in of the alignment on hydrogen bond donors and acceptors.

We did not observe any cases of CSMs for which their targets were not ranked early in the hit list and at least one known ligand shared a high degree of 2D similarity with the

Table 6. Examples of CSMs for Which Their Targets Were Successfully Identified Despite Being Dissimilar from Any Reference Compound

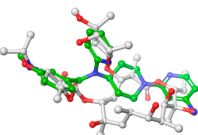
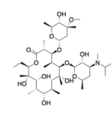
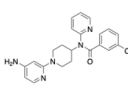
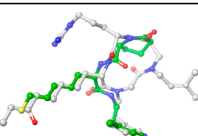
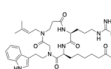
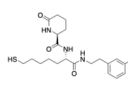
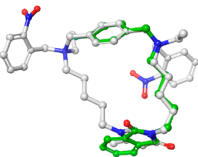
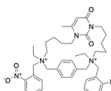
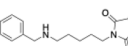
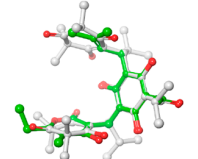
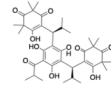
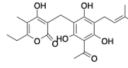
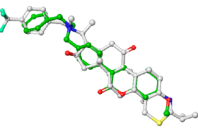
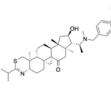
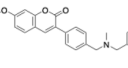
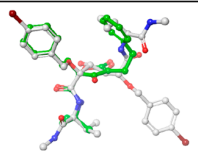
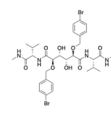
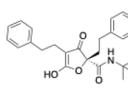
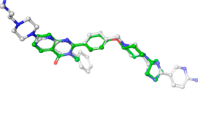
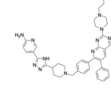
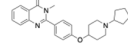
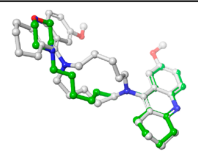
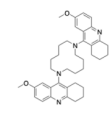
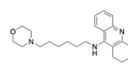
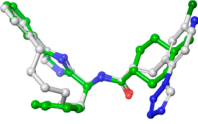
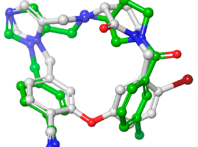
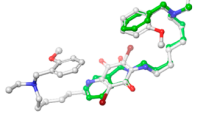
Query ^a	Closest Reference compound	3D alignment	2D similarity ^b	Tanimoto Combo score	Target rank with Tanimoto Combo	Target ^c
CHEMBL584549*	CHEMBL493517		0.08	0.71	7	HERG
						
CHEMBL2170017*	CHEMBL3356937		0.12	0.73	1	HDAC1
						
CHEMBL3621333*	CHEMBL3415568		0.12	0.77	1	AChE
						
CHEMBL508629	CHEMBL225421		0.13	1.09	1	PGES
						
CHEMBL1783518	CHEMBL413793		0.13	1.05	5	AChE
						
CHEMBL124139	CHEMBL104253		0.15	0.78	5	HIV-1 protease
						
CHEMBL503270	CHEMBL455681		0.15	0.83	1	HERG
						
CHEMBL1917826*	CHEMBL1819169		0.16	1.11	1	AChE
						

Table 6. continued

Query ^a	Closest Reference compound	3D alignment	2D similarity ^b	Tanimoto Combo score	Target rank with Tanimoto Combo	Target ^c
CHEMBL3676156*	CHEMBL3393362; CH EMBL3355664		0.17	1.07	3	F11
CHEMBL524997*	CHEMBL317520		0.18	1.30	1	HERG
CHEMBL243062	CHEMBL3402709		0.18	0.79	8	AChE

^aQueries marked with a “*” are macrocyclic compounds. ^b2D molecular similarity between the CSM query and the closest ligand recorded in the knowledge base (measured as Tanimoto coefficient based on Morgan2 fingerprints). ^cHDAC1, histone deacetylase 1; AChE, acetylcholinesterase; PGES, prostaglandin E synthase; HIV-1 protease, human immunodeficiency virus type 1 protease; F11, coagulation factor XI.

query (note that the number of CSMs in this category was small).

Performance as a Function of Common Substructures. Target rankings are expected to improve with the size of the maximum common substructure (MCS) shared between the CSM query and the closest related non-CSM in the knowledge base (as determined by ROCS). The results presented in Figure 12 confirm this assumption: For the TanimotoCombo score, the median ranking of the targets of interest was 3.5 for CSMs sharing an MCS of at least 20 heavy atoms with the closest ligand (non-CSM) recorded in the knowledge base, whereas the median target rank was just 111.5 for CSMs with an MCS of 15 to 19 heavy atoms. The median target ranks obtained by the RefTverskyCombo, FitTversky-Combo, and ShapeTanimoto scores were substantially lower (worse): 28, 80, and 43 for CSMs sharing an MCS of a least 20 heavy atoms, respectively, and 318, 299, and 227 for CSMs with an MCS of 15 to 19 heavy atoms, respectively. We repeated this analysis using the percentage of heavy atoms rather than absolute numbers covered by the MCSs and observed the same trends (data not shown).

Performance on Natural Products. By overlapping the queries with a data set of 201 761 natural products compiled as part of the work reported in ref 41, we determined that at least six out of the 269 (unique) CSMs are natural products (which is a surprisingly low portion of natural products). We employed NP-Scout⁴¹ to identify additional CSMs that likely are natural products or natural product-like. NP-Scout is a random forest classifier discriminating between natural products and synthetic molecules. The model is trained on 108 393 natural products and 157 162 synthetic molecules

represented by MACCS keys. The model yielded an AUC of 0.997 and Matthews correlation coefficient of 0.960 during tests with external data. NP-Scout identified an additional 20 CSMs with a high likelihood (probability >0.70) of being natural products.

The 26 natural products and natural product-like compounds cover a total of 18 different targets; eight of the queries are macrocyclic. Using the TanimotoCombo score, ROCS ranked the targets of interest of the natural products among the top-10 positions for only seven out of 31 queries (23%; the 31 queries result from the 26 unique natural products and natural product-like compounds). This success rate is considerably lower than the ones averaged over all 280 queries (37%), all 245 nonmacrocyclic queries (39%), and all macrocyclic queries (27%), indicating that the prediction of the targets of complex natural products is more challenging than of complex synthetic molecules. A main reason for the low prediction success rates is the fact that the similarity of complex natural products and natural product-like compounds and the nearest non-CSMs of the knowledge base is generally low: The median Tanimoto coefficient based on Morgan2 fingerprints for these types of CSMs and the non-CSMs of the knowledge based is only 0.13, whereas it is 0.21 for the other CSMs and their closest non-CSMs).

Runtimes. The ROCS screening process takes less than 6 h per CSM query on a single core of an i5-4590 CPU at 3.30 GHz. Runtimes are therefore expected not to pose a barrier to the usability of the method.

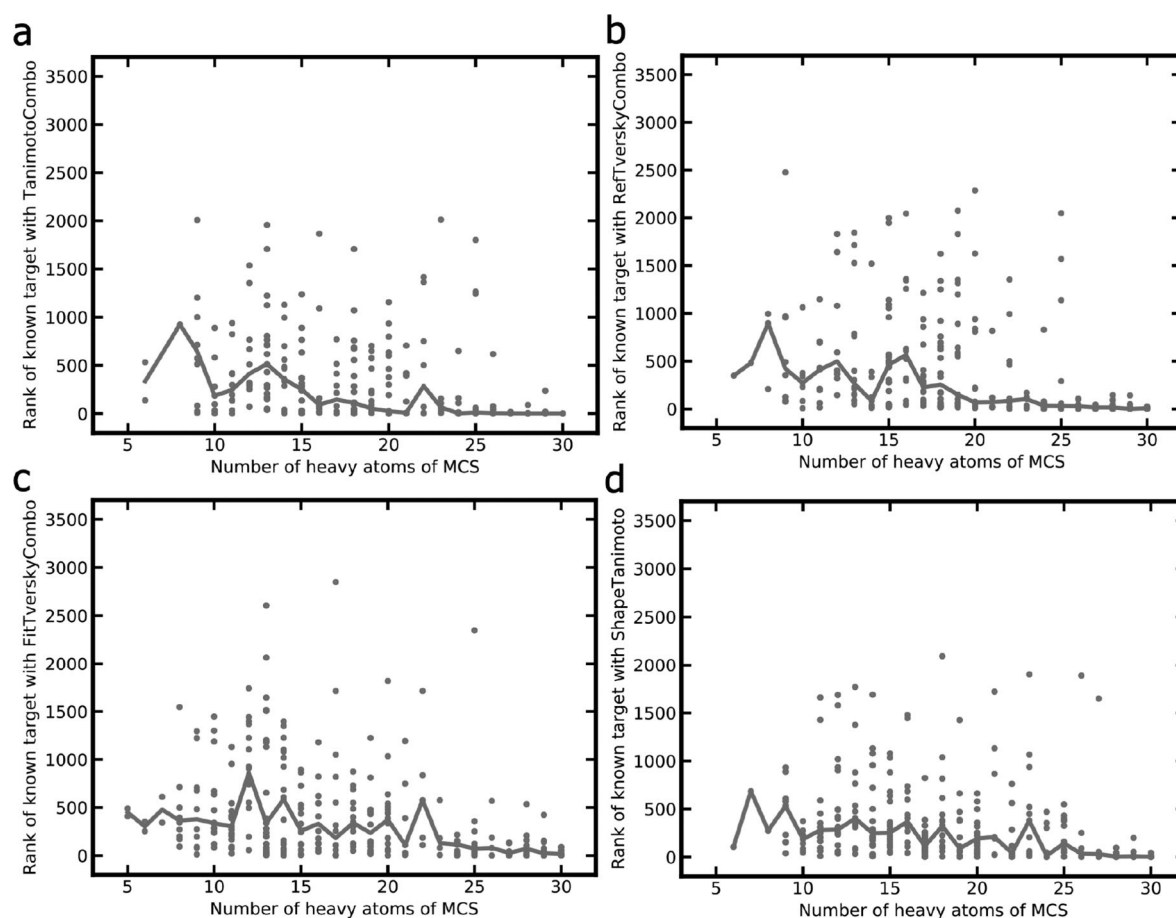


Figure 12. Ranks obtained for the targets of interest as a function of the size of the MCS shared between the CSM queries and most similar ligand (non-CSM) recorded for the respective target for the (a) TanimotoCombo, (b) RefTverskyCombo, (c) FitTverskyCombo, and (d) ShapeTanimoto scores. The lines are merely a guide for the eye and indicate the median values of the target rankings in relation to the size of the MCS.

CONCLUSIONS

In this work, we showed that the 3D alignment-dependent shape-based methods ROCS, in combination with the best-performing scoring function, the TanimotoCombo score, ranks the targets of approximately one-third of 280 investigated CSM queries among the top-5 ranks of hit lists of more than 3600 proteins. The success rate increases to 41% if the top-20 ranks are considered. For 24 of the 28 proteins (86%), the target of interest was ranked at the top position with at least one of the 10 queries. These results indicate that the method may well be a valuable tool for prioritizing research efforts in early drug discovery because researchers, with their expert knowledge and background information on a compound of interest (e.g., observations from phenotypic assays), will likely be able to rule out many of the proteins wrongly predicted as targets.

An important advantage of ROCS is its use of hard Gaussians for describing chemical features (color), which causes a lock-in effect during alignment. Alignments produced by ROCS therefore typically look “tidy”, enabling chemists to easily interpret the results and make their own judgements on the reliability of individual predictions (thereby excluding many false-positive predictions). Even if none of the predictions are deemed plausible, e.g., because of the lack of any good matches with compounds in the knowledge base, this

can be valuable information as it is a good indication for a compound being novel and perhaps targeting a so-far unexplored biomacromolecule (or having a distinct binding mode). An important advantage of similarity-based approaches over many other methods is that the final prediction relies on a single data point (as opposed to, for example, machine learning approaches), making it straightforward for researchers to verify the reliability of that specific data point with the primary literature data.

Also, for 3D alignment-dependent shape-based methods, the success rates for the prediction of the targets of CSMs decline with decreasing molecular similarity between the CSM query and the ligands in the knowledge base. Macrocyclic compounds and natural products prove to be particularly challenging to the approach. Nevertheless, the robustness of the approach is impressive, given the fact that structurally highly dissimilar molecules, even though binding to the same binding site, may likely exhibit distinct binding modes, which is beyond the reach of any ligand-based approach.

Taking performance, usability, and interpretability into account, we believe that 3D alignment-dependent shape-based approaches such as the one investigated in this work are predestined for use in target prediction for CSMs and molecules for which data on structurally related compounds are scarce. With the increasing amount of bioactivity data, the

reach and value of these and related methods will continue to improve.

DATA AVAILABILITY

The complete sets of CSMs and non-CSMs (including the original SMILES notations from ChEMBL, ChEMBL compound IDs, natural product-likeness scores, and labels for macrocycles) are available on GitHub at https://github.com/anya-chen/CSMs_target_prediction.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.0c00161>.

Density distribution of the molecular weight and number of heavy atoms of compounds of the processed data set and Approved Drugs subset of DrugBank (PDF)

AUTHOR INFORMATION

Corresponding Author

Johannes Kirchmair – Center for Bioinformatics (ZBH), Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany; Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway; Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria; orcid.org/0000-0003-2667-5877; Phone: +43-1-4277-55104; Email: johannes.kirchmair@univie.ac.at

Authors

Ya Chen – Center for Bioinformatics (ZBH), Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany; orcid.org/0000-0001-5273-1815

Neann Mathai – Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway; orcid.org/0000-0002-5763-6304

Complete contact information is available at <https://pubs.acs.org/doi/10.1021/acs.jcim.0c00161>

Funding

Y.C. is supported by the China Scholarship Council (201606010345). N.M. and J.K. are supported by the Trond Mohn Foundation (BFS2017TMT01).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Christina de Bruyn Kops from the Universität Hamburg and Christoph Bauer from the University of Bergen for valuable discussions and OpenEye for providing an academic license for the use of OMEGA and ROCS.

ABBREVIATIONS

CSM, complex small molecule; MCS, maximum common substructure

REFERENCES

(1) Sydow, D.; Burggraaff, L.; Szengel, A.; van Vlijmen, H. W. T.; Ijzerman, A. P.; van Westen, G. J. P.; Volkamer, A. *Advances and*

Challenges in Computational Target Prediction. J. Chem. Inf. Model. **2019**, *59*, 1728–1742.

(2) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. Tools for in silico Target Fishing. *Methods* **2015**, *71*, 98–103.

(3) Agamah, F. E.; Mazandu, G. K.; Hassan, R.; Bope, C. D.; Thomford, N. E.; Ghansah, A.; Chimusa, E. R. Computational/in silico Methods in Drug Target and Lead Prediction. *Briefings Bioinf.* **2019**, DOI: 10.1093/bib/bbz103.

(4) Moffat, J. G.; Vincent, F.; Lee, J. A.; Eder, J.; Prunotto, M. Opportunities and Challenges in Phenotypic Drug Discovery: An Industry Perspective. *Nat. Rev. Drug Discovery* **2017**, *16*, 531–543.

(5) Mathai, N.; Chen, Y.; Kirchmair, J. Validation Strategies for Target Prediction Methods. *Briefings Bioinf.* **2019**, DOI: 10.1093/bib/bbz026.

(6) Wang, C.; Kurgan, L. Review and Comparative Assessment of Similarity-Based Methods for Prediction of Drug-Protein Interactions in the Druggable Human Proteome. *Briefings Bioinf.* **2019**, *20*, 2066–2087.

(7) Lo, Y.-C.; Senese, S.; Li, C.-M.; Hu, Q.; Huang, Y.; Damoiseaux, R.; Torres, J. Z. Large-Scale Chemical Similarity Networks for Target Profiling of Compounds Identified in Cell-Based Chemical Screens. *PLoS Comput. Biol.* **2015**, *11*, No. e1004153.

(8) Boezio, B.; Audouze, K.; Ducrot, P.; Taboureaux, O. Network-Based Approaches in Pharmacology. *Mol. Inf.* **2017**, *36*, 1700048.

(9) Rodrigues, T.; Bernardes, G. J. L. Machine Learning for Target Discovery in Drug Development. *Curr. Opin. Chem. Biol.* **2020**, *56*, 16–22.

(10) Sam, E.; Athri, P. Web-Based Drug Repurposing Tools: A Survey. *Briefings Bioinf.* **2019**, *20*, 299–316.

(11) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47*, D1102–D1109.

(12) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Motow, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, S.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.

(13) PubChem. <https://pubchem.ncbi.nlm.nih.gov/> (accessed Feb 4, 2020).

(14) ChEMBL Database. <https://www.ebi.ac.uk/chembl/> (accessed Apr 6, 2020).

(15) Moumbock, A. F. A.; Li, J.; Mishra, P.; Gao, M.; Günther, S. Current Computational Methods for Predicting Protein Interactions of Natural Products. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 1367–1376.

(16) Cockroft, N. T.; Cheng, X.; Fuchs, J. R. S*StarFish: A Stacked Ensemble Target Fishing Approach and Its Application to Natural Products. *J. Chem. Inf. Model.* **2019**, *59*, 4906–4920.

(17) Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* **2017**, *57*, 2099–2111.

(18) Marsault, E.; Peterson, M. L. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *J. Med. Chem.* **2011**, *54*, 1961–2004.

(19) You, L.; An, R.; Liang, K.; Cui, B.; Wang, X. Macrocyclic Compounds: Emerging Opportunities for Current Drug Discovery. *Curr. Pharm. Des.* **2016**, *22*, 4086–4093.

(20) Mallinson, J.; Collins, I. Macrocycles in New Drug Discovery. *Future Med. Chem.* **2012**, *4*, 1409–1438.

(21) Reker, D.; Perna, A. M.; Rodrigues, T.; Schneider, P.; Reutlinger, M.; Mönch, B.; Koeberle, A.; Lamers, C.; Gabler, M.; Steinmetz, H.; Müller, R.; Schubert-Zsilavecz, M.; Werz, O.; Schneider, G. Revealing the Macromolecular Targets of Complex Natural Products. *Nat. Chem.* **2014**, *6*, 1072–1078.

(22) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How To Optimize Shape-Based Virtual

Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **2009**, *49*, 678–692.

(23) Gfeller, D.; Michielin, O.; Zoete, V. Shaping the Interaction Landscape of Bioactive Molecules. *Bioinformatics* **2013**, *29*, 3073–3079.

(24) Potshangbam, A. M.; Polavarapu, R.; Rathore, R. S.; Naresh, D.; Prabhu, N. P.; Potshangbam, N.; Kumar, P.; Vindal, V. MedPServer: A Database for Identification of Therapeutic Targets and Novel Leads Pertaining to Natural Products. *Chem. Biol. Drug Des.* **2019**, *93*, 438–446.

(25) Gong, J.; Cai, C.; Liu, X.; Ku, X.; Jiang, H.; Gao, D.; Li, H. ChemMapper: A Versatile Web Server for Exploring Pharmacology and Chemical Structure Association Based on Molecular 3D Similarity Method. *Bioinformatics* **2013**, *29*, 1827–1829.

(26) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping[®] by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.

(27) Kumar, A.; Zhang, K. Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front. Chem. (Lausanne, Switz.)* **2018**, *6*, 315.

(28) ChEMBL Database, version 25. <https://www.ebi.ac.uk/chembl/> (accessed May 15, 2019).

(29) Bosc, N.; Atkinson, F.; Felix, E.; Gaulton, A.; Hersey, A.; Leach, A. R. Large Scale Comparison of QSAR and Conformal Prediction Methods and Their Applications in Drug Discovery. *J. Cheminf.* **2019**, *11*, 4.

(30) Molecular Operating Environment (MOE). Chemical Computing Group. <https://www.chemcomp.com/Products.htm>.

(31) Hawkins, P. C. D.; Nicholls, A. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model.* **2012**, *52* (11), 2919–2936.

(32) OMEGA 3.1.1.2. OpenEye Scientific Software. <https://www.eyesopen.com> (accessed Nov 13, 2019).

(33) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 2719–2728.

(34) Poongavanam, V.; Danelius, E.; Peintner, S.; Alcaraz, L.; Caron, G.; Cummings, M. D.; Wlodek, S.; Erdelyi, M.; Hawkins, P. C. D.; Ermondi, G.; Kihlberg, J. Conformational Sampling of Macrocyclic Drugs in Different Environments: Can We Find the Relevant Conformations? *ACS Omega* **2018**, *3*, 11742–11757.

(35) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28*, 3150–3152.

(36) CD-HIT Suite. <http://weizhong-lab.ucsd.edu/cdhit-web-server/cgi-bin/index.cgi?cmd=cd-hit> (accessed Mar 17, 2020).

(37) ROCS 3.3.1.2. OpenEye Scientific Software. <https://www.eyesopen.com> (accessed Nov 13, 2019).

(38) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74–82.

(39) Méndez-Lucio, O.; Medina-Franco, J. L. The Many Roles of Molecular Complexity in Drug Discovery. *Drug Discovery Today* **2017**, *22*, 120–126.

(40) Rasina, D.; Otkovs, M.; Leitans, J.; Recacha, R.; Borysov, O. V.; Kanepe-Lapsa, I.; Domraceva, I.; Pantelejevs, T.; Tars, K.; Blackman, M. J.; Jaudzems, K.; Jirgensons, A. Fragment-Based Discovery of 2-Aminoquinazolin-4(3H)-Ones As Novel Class Non-peptidomimetic Inhibitors of the Plasmepsins I, II, and IV. *J. Med. Chem.* **2016**, *59*, 374–387.

(41) Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **2019**, *9*, 43.

4. Concluding Discussion

In recent years, the amount of data available on the chemical, biological, pharmacological and structural properties of NPs has increased dramatically. This has fueled the development and application of cheminformatics methods in the context of NP research. However, the quantity, quality and relevance of the available data on NPs are poorly understood, and the scope and limitations of cheminformatics methods often undefined in the context of NP research.

Starting from a comprehensive literature survey of current applications of cheminformatics methods in NP research ([Chapter 1.1](#)) and an exhaustive analysis of NP data resources relevant to cheminformatics ([Chapter 2.1](#)), we conducted a detailed and thorough characterization of the physicochemical and structural properties of natural products, as well as the chemical space covered by different NP databases ([Chapter 3.1](#)). Utilizing the collected chemical data on pure NPs, we developed NP-Scout, a machine learning approach for identifying natural products and natural product-like compounds ([Chapter 3.2](#)). NP-Scout features a visualizer that highlights atoms in a molecule which contribute to the classification of a compound as a natural product or as a synthetic compound. In the last part of this work, the ability of a 3D shape-based method for predicting biomacromolecular targets of structurally CSMs, including NPs, was determined ([Chapter 3.3](#)).

From our comprehensive analysis of the existing physical and virtual NP databases ([Chapter 2.1](#)) we learned that approximately 250k NPs are known to date, of which roughly 10% (25k) are readily obtainable for testing. An additional 10k to 30k of readily obtainable, NP-like compounds were identified using a 2D similarity search. Important lessons learned from this large-scale data analysis are that few of the databases are sustainably maintained and that the stereochemical information provided is often incomplete and sometimes even wrong. In order to obtain a more detailed and clean picture of the content of the individual NP data sets and characterize the physicochemical property space of the represented NPs, we devised a method for the automated removal of sugars and sugar-like moieties from NPs ([Chapter 3.1](#)) as these moieties are rarely part of the pharmacophore and as such not essential for bioactivity.

Most of the previous studies on NPs analyzed molecular structure focusing on cyclic systems, overlooking some biologically meaningful scaffolds of NPs that are not ring systems or are the combinations of ring systems and linkers. In our work ([Chapter 3.1](#)), we devised an automated approach based on SMARTS patterns for the classification of NPs into the major NP classes, including alkaloids, flavonoids (as well as subclasses of flavonoids), and steroids. With this rule set we found that some NP databases are particularly rich in certain NP classes. For example, StreptomeDB 2.0 [38] has a high proportion of alkaloids (47%), and the South African Natural Compounds Database (SANCDB) [39,40] has the highest rates of steroids (14%) compared to other virtual NP databases. The rule set was also employed in a study (A2) investigating the capacity of machine learning methods to identify compounds that have a higher than expected hit rate in biological

assays (frequent hitters). Note that we used strict definitions for the subclasses of flavonoids in our study, and that further development could extend the defined patterns by matching more exceptions and covering more subclasses of flavonoids, as well as classification of other types of NPs. Because of many repeated units and exceptions, for some classes, algorithms which employ rules beyond just using SMARTS patterns for structure detection may be also worth exploring.

As we also learned from our study, the known NPs cover a much wider chemical space than approved drugs, and a large number of NPs populate areas of the chemical space that are covered by approved drugs. This explains why NPs are one of the most prolific sources of inspiration for drug discovery. In particular the readily obtainable NPs are highly diverse, representing more than 5700 different Murcko scaffolds and covering all of the major NP classes. Readily obtainable NPs are also highly relevant to the chemical space covered by approved drugs and around two-thirds of them are fragment-sized thus some of them could serve as good start points for optimization. Of relevance to structure-based drug design, we identified high quality X-ray crystal structures of more than 2000 different NPs bound to at least one biomacromolecule in the PDB. These NPs are generally smaller-sized and more hydrophilic than approved drugs.

Some distinctive features of individual databases were also identified. For example, the NPs subset of the PubChem Substance Database [41,42], which contains NPs and their associated bioactivity data, stands out due to its high proportion of drug-like NPs, and the Traditional Chinese Medicine Database@Taiwan [43,44], the largest freely available source of traditional Chinese medicine data, is characterized by the coverage of a wide and in part unique chemical space containing many large and highly chiral NPs.

As newer data sources of NPs become available, it would be interesting to re-evaluate the quality, quantity and availability of NPs. Despite the challenges, which include data availability, quality and sustainability, computational methods will be a key contributor to NP research and NPs will continue to inform drug discovery.

Among the data quality issues of NP databases, is that many of them have NPs mixed with NP derivatives and analogs even though they claim to provide only genuine NPs. Also, many synthetic compound libraries contain a significant number of NPs which are not labeled or explicitly mentioned. These data challenges were our motivation for the development of NP-Scout: a machine learning approach to identify and visualize NPs and NP-like compounds (Chapter 3.2). NP-Scout is built of random forest classification models which were trained on a large collection of pure NPs and an equal number of synthetic molecules. From principal component analysis, based on the main physicochemical properties, we found that although presented by an equal number of unique structures, the NPs cover a much larger chemical space than the synthetic molecules. There are also some clear differences in individual physicochemical properties. For example, NPs have, on average, a higher molecular weight and lower element distribution entropy than synthetic molecules. NPs also tend to have more chiral centers, fewer nitrogen atoms and more oxygen atoms.

All three NP-Scout models, based on three different sets of molecular descriptors or fingerprints, performed very well. The models achieved AUCs of

0.997 and MCCs of 0.954 and higher on the test set. The best performing model was able to predict approximately 95% of compounds in the Dictionary of Natural Products not represented in the training set as NPs. The model performs similar to an earlier well-known method [45] based on Bayesian statistics. We used a Java implementation of this method called NP-Likeness calculator [46,47] that allows the use of customized data sets for training. When using the same training set and test set as ours, the NP-Likeness calculator performed comparable to our models, with an AUC of 0.997 and an MCC of 0.959. Furthermore, the applications of our best model to the ChEMBL database [48,49] and other two datasets (the ChEMBL subset of molecules published in the *Journal of Natural Products* and the NPs subset of ZINC database) showed its ability to distinguish NPs and synthetic molecules, especially indicating the existence of synthetic molecules in these two datasets which are often used as libraries of genuine NPs.

To understand and interpret the workings of these models, we first analyzed the important features contributing to the classifications. For the classifier based on MOE 2D molecular descriptors, the three most important features were the number of nitrogen atoms, the entropy of the element distribution in molecules and the number of unconstrained chiral centers, whose differences were already directly or indirectly seen in the physicochemical properties analysis. For the most relevant MACCS keys, also in agreement with the difference in the physicochemical properties, the most important key describes the presence or absence of nitrogen atoms and other important keys involve substructures with nitrogen and/or oxygen atoms. Moreover, the utilization of similarity maps [50] allows the visualization of atoms of a molecule which are characteristic to NPs or synthetic compounds, according to the Morgan2 fingerprint-based model.

The models are accessible as a free web service at <https://nerdd.zbh.uni-hamburg.de/npscout/>. The web service returns the NP class probability of the given molecules and shows the similarity maps, highlighting the NP-like or synthetic-like fragments in the molecules. The method can therefore be utilized to cherry-pick NPs and NP-like compounds from large molecular libraries, quantify the NP-likeness of small molecules, and visualize the atoms in small molecules which contribute to the classification of NPs or synthetic compounds. This model has also been used in characterization of several other datasets, including a comprehensive set of small-molecule ligands observed in high-quality co-crystals in the PDB (A3), the “in-stock” subset of ZINC database (A3), as well as used in the last part of this dissertation (Chapter 3.3, D6). Additionally, NP-Scout is currently being used in several ongoing virtual screening campaigns. Here, the NP class probabilities for compounds from multiple commercial screening databases are being used to identify bioactive NP-like compounds of interest.

Many NPs have distinct molecular structures and physicochemical properties. The fact that data on NPs are scarce makes target prediction a difficult task for NPs, in particular for complex NPs. Methods utilizing 3D molecular shape representations for the comparison of molecules may be able to recognize relationships between compounds which are less similar. Hence, such methods are predestined for use with CSMs such as many NPs. Therefore, in the last part of this thesis (Chapter 3.3), we systematically investigated the capacity of ROCS, a

leading shaped-based approach, to identify the macromolecular targets of CSMs from non-CSMs.

The definition of molecular complexity depends on the context and there is no universally applicable and easily interpretable metric for its quantification [51]. It should be noted that different complexity measures capture different aspects and a metric that is defined simply is not necessarily a poor measure of complexity, but may in fact be effective at quantifying molecular complexity while being easily interpretable. For the purpose of this work, we defined molecules as "complex" if they are either very large in size (45 to 55 heavy atoms) or macrocyclic (and large). In contrast, we defined molecules as "non-complex" if they were small in size (15 to 30 heavy atoms).

A total of 28 pharmaceutically relevant targets were studied and for each target a diverse set of 10 CSMs was generated. Using a knowledge base of non-complex compounds with measured bioactivity data, a retrospective study to predict the targets of 280 CSM queries was conducted. Approximately one-third of these queries had the known target ranked among the top-5 of the possible 3642 proteins when the best-performing scoring function, the TanimotoCombo score, was used. The success rate increases to 41% if the top-20 ranks are considered. For 24 of the 28 proteins (86%), the target of interest was ranked at the top position with at least one of the ten queries. These results indicate that the method may be valuable for prioritizing research efforts in early drug discovery. Using the predictions, researchers will likely be able to rule out many of the proteins wrongly predicted as targets based on their expert knowledge and background information on a compound of interest (e.g. observations from phenotypic assays).

ROCS uses hard Gaussians to describe chemical features, so the alignments are easy to interpret. Researchers can therefore make their own decision on the reliability of the individual predictions based on these alignments, and exclude many false-positive predictions. When there is no plausible prediction, the results can indicate the novelty of the compound, e.g. targeting unexplored targets or having distinct binding mode. In general, for similarity-based approaches, the final predictions are based on individual data points that are straightforward to verify from primary sources such as literature reports.

At least 31 known, complex NPs and NP-like compounds (identified by our collection of NPs and NP-Scout) were among the 280 CSMs. For these compounds, the success rate was lower. For example, when the top-10 ranks were considered the success rate for these queries is only 23% vs. 37% for all queries. This is related to the fact that the median Tanimoto coefficient based on Morgan2 fingerprints of the complex NP or NP-like compound and the closest non-complex small molecule in the knowledge base is only 0.13. The success rates for the prediction of the targets of CSMs decline with decreasing 2D molecular similarity between the CSM query and the closest compound in the knowledge base. For pairs of compounds sharing such low degree of similarity their binding modes are likely to be distinct, which is generally beyond the scope of ligand-based methods.

The similarity of a CSM query to its closest compound in the knowledge base was measured in different ways in order to understand how performance varied with this relationship. The Tanimoto coefficient, using both Morgan2 fingerprints and atom types fingerprints, were used to quantify the relationship between a

query and the knowledge base. With both fingerprints, the success rates declined as the Tanimoto coefficients decreased. Also the Tversky coefficient, which measures the asymmetric similarity between two compounds, showed the same trends as the Tanimoto coefficient.

Taking performance, usability and interpretability into account, we believe that 3D alignment-dependent, shape-based approaches such as the one we investigated are capable of predicting targets for molecules for which data on structurally related compounds are scarce. With the increasing amount of bioactivity data, the reach and value of these and related methods will continue to improve.

Overall, the studies presented in this thesis have resulted in a clean and comprehensive picture of the quantity, quality and availability on the data and cheminformatics methods relevant to natural products-based drug discovery. The work has also resulted in machine learning models which are able to discriminate between NPs and synthetic molecules with high accuracy. These models have garnered significant interest from the scientific community and are frequently accessed via the free web service. Moreover, we have determined the scope and limitations of a 3D shape-based approach for the prediction of the biomacromolecular targets of CSMs including NPs. We expect that data on more NP compounds, richer annotations on NP biological activities and ADME properties, and NP data with special focuses will be added to the public domain in the future. The growing volume of data and interest in NPs is coupled with the results of our studies, which confirm the relevance of NPs as an important source of inspiration in drug discovery. We therefore believe that the insights gained through the research presented in this dissertation will help with data selection, method development and future NP research.

Bibliography

- 1 Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83* (3), 770–803.
- 2 Cragg, G. M.; Newman, D. J. Biodiversity: A Continuing Source of Novel Drug Leads. *Pure Appl. Chem.* **2005**, *77* (1), 7–24.
- 3 Rodrigues, T.; Reker, D.; Schneider, P.; Schneider, G. Counting on Natural Products for Drug Design. *Nat. Chem.* **2016**, *8* (6), 531–541.
- 4 Atanasov, A. G.; Waltenberger, B.; Pferschy-Wenzig, E.-M.; Linder, T.; Wawrosch, C.; Uhrin, P.; Temml, V.; Wang, L.; Schwaiger, S.; Heiss, E. H.; Rollinger, J. M.; Schuster, D.; Breuss, J. M.; Bochkov, V.; Mihovilovic, M. D.; Kopp, B.; Bauer, R.; Dirsch, V. M.; Stuppner, H. Discovery and Resupply of Pharmacologically Active Plant-Derived Natural Products: A Review. *Biotechnol. Adv.* **2015**, *33* (8), 1582–1614.
- 5 Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PLoS One* **2013**, *8*, e62839.
- 6 Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small Molecules of Different Origins Have Distinct Distributions of Structural Complexity That Correlate with Protein-Binding Profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (44), 18787–18792.
- 7 Chen, H.; Engkvist, O.; Blomberg, N.; Li, J. A Comparative Analysis of the Molecular Topologies for Drugs, Clinical Candidates, Natural Products, Human Metabolites and General Bioactive Compounds. *Med. Chem. Commun.* **2012**, *3*, 312–321.
- 8 David, B.; Grondin, A.; Schambel, P.; Vitorino, M.; Zeyer, D. Plant Natural Fragments, an Innovative Approach for Drug Discovery. *Phytochem. Rev.* **2019**. <https://doi.org/10.1007/s11101-019-09612-4>.
- 9 Olğaç, A.; Orhan, I. E.; Banoglu, E. The Potential Role of in silico Approaches to Identify Novel Bioactive Molecules from Natural Resources. *Future Med. Chem.* **2017**, *9* (14), 1665–1686.
- 10 Rodrigues, T. Harnessing the Potential of Natural Products in Drug Discovery from a Cheminformatics Vantage Point. *Org. Biomol. Chem.* **2017**, *15* (44), 9275–9282.
- 11 Ikram, N. K. K.; Durrant, J. D.; Muchtaridi, M.; Zalaludin, A. S.; Purwitasari, N.; Mohamed, N.; Rahim, A. S. A.; Lam, C. K.; Normi, Y. M.; Rahman, N. A.;

- Amaro, R. E.; Wahab, H. A. A Virtual Screening Approach for Identifying Plants with Anti H₅N₁ Neuraminidase Activity. *J. Chem. Inf. Model.* **2015**, *55* (2), 308–316.
- 12 Grienke, U.; Mihály-Bison, J.; Schuster, D.; Afonyushkin, T.; Binder, M.; Guan, S.-H.; Cheng, C.-R.; Wolber, G.; Stuppner, H.; Guo, D.-A.; Bochkov, V. N.; Rollinger, J. M. Pharmacophore-Based Discovery of FXR-Agonists. Part II: Identification of Bioactive Triterpenes from *Ganoderma Lucidum*. *Bioorg. Med. Chem.* **2011**, *19* (22), 6779–6791.
- 13 Rollinger, J. M.; Kratschmar, D. V.; Schuster, D.; Pfisterer, P. H.; Gumy, C.; Aubry, E. M.; Brandstötter, S.; Stuppner, H.; Wolber, G.; Odermatt, A. α -Hydroxysteroid Dehydrogenase 1 Inhibiting Constituents from *Eriobotrya Japonica* Revealed by Bioactivity-Guided Isolation and Computational Approaches. *Bioorg. Med. Chem.* **2010**, *18* (4), 1507–1515.
- 14 Grienke, U.; Braun, H.; Seidel, N.; Kirchmair, J.; Richter, M.; Krumbholz, A.; von Grafenstein, S.; Liedl, K. R.; Schmidtke, M.; Rollinger, J. M. Computer-Guided Approach to Access the Anti-Influenza Activity of Licorice Constituents. *J. Nat. Prod.* **2014**, *77* (3), 563–570.
- 15 Berthold, M. R.; Cebon, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization*; Baier, D., Critchley, F., Decker, R., Diday, E., Greenacre, M., Lauro, C.N., Meulman, J., Monari, P., Nishisato, S., Ohsumi, N., Opitz, O., Ritter, G., Schader, M., Ed.; Springer: Berlin, 2007; pp 319–326.
- 16 Molecular Operating Environment (MOE), Version 2016.08; Chemical Computing Group, Montreal, QC.
- 17 RDKit: Open-Source Cheminformatics Software. <http://www.rdkit.org> (accessed May 6, 2016).
- 18 Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31–36.
- 19 Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7*, 23.
- 20 InChI, Version 1.05; IUPAC: Research Triangle Park, NC, 2017.
- 21 Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> (accessed Feb 6, 2017).
- 22 Hänsel, R.; Sticher, O. *Pharmakognosie - Phytopharmazie*; Springer Berlin Heidelberg, 2009.
- 23 Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion,

- B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- 24 Scikit-Learn: Machine Learning in Python, Version 0.19.1.
- 25 Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.
- 26 Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- 27 Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta* **1975**, *405* (2), 442–451.
- 28 Dictionary of Natural Products, Version 19.1; Chapman & Hall/CRC, 2010.
- 29 ROCS 3.3.1.2. OpenEye Scientific Software. <https://www.eyesopen.com> (accessed Nov 13, 2019).
- 30 Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50* (1), 74–82.
- 31 Singh, S. B.; Culberson, C. J. Chemical Space and the Difference Between Natural Products and Synthetics. In *Natural Product Chemistry for Drug Discovery*; Buss, A. D., Butler, M. S., Eds.; The Royal Society of Chemistry: Cambridge, United Kingdom, 2009; pp 28–43.
- 32 Grabowski, K.; Baringhaus, K.-H.; Schneider, G. Scaffold Diversity of Natural Products: Inspiration for Combinatorial Library Design. *Nat. Prod. Rep.* **2008**, *25*, 892–904.
- 33 Ertl, P.; Schuffenhauer, A. Cheminformatics Analysis of Natural Products: Lessons from Nature Inspiring the Design of New Drugs. *Prog. Drug Res.* **2008**, *66*, 217, 219–235.
- 34 Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337.
- 35 ZINC15. <http://zinc15.docking.org> (accessed May 26, 2017).
- 36 Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- 37 RCSB Protein Data Bank. <https://www.rcsb.org> (accessed Feb 12, 2018).
- 38 Klementz, D.; Döring, K.; Lucas, X.; Telukunta, K. K.; Erxleben, A.; Deubel, D.; Erber, A.; Santillana, I.; Thomas, O. S.; Bechthold, A.; Günther, S.

- StreptomeDB 2.0—an Extended Resource of Natural Products Produced by Streptomyces. *Nucleic Acids Research*. **2016**, *44* (D1), D509–D514.
- 39 Hatherley, R.; Brown, D. K.; Musyoka, T. M.; Penkler, D. L.; Faya, N.; Lobb, K. A.; Tastan Bishop, Ö. SANCDB: A South African Natural Compound Database. *J. Cheminf.* **2015**, *7*, 29.
- 40 South African Natural Compound Database (SANCDB). <http://sancdb.rubi.ru.ac.za> (accessed Feb 8, 2017).
- 41 Hao, M.; Cheng, T.; Wang, Y.; Bryant, H. S. Web Search and Data Mining of Natural Products and Their Bioactivities in PubChem. *Sci. China Chem.* **2013**, *56*, 1424–1435.
- 42 PubChem Substance. <http://ncbi.nlm.nih.gov/pcsubstance> (accessed Apr 7, 2017).
- 43 Chen, C. Y.-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening in silico. *PLoS One* **2011**, *6*, e15939.
- 44 TCM Database@Taiwan. <http://tcm.cmu.edu.tw> (accessed Oct 17, 2016).
- 45 Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-Likeness Score and Its Application for Prioritization of Compound Libraries. *J. Chem. Inf. Model.* **2008**, *48* (1), 68–74.
- 46 Jayaseelan, K. V.; Moreno, P.; Truszkowski, A.; Ertl, P.; Steinbeck, C. Natural Product-Likeness Score Revisited: An Open-Source, Open-Data Implementation. *BMC Bioinformatics* **2012**, *13*, 106.
- 47 Natural Product Likeness Calculator Version 2.1. <https://sourceforge.net/projects/np-likeness/> (accessed Oct 5, 2018).
- 48 Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.
- 49 ChEMBL Version 24_1. <https://www.ebi.ac.uk/chembl/> (accessed Jul 30, 2018).
- 50 Riniker, S.; Landrum, G. A. Similarity Maps - a Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J. Cheminf.* **2013**, *5* (1), 43.
- 51 Méndez-Lucio, O.; Medina-Franco, J. L. The Many Roles of Molecular Complexity in Drug Discovery. *Drug Discov. Today* **2017**, *22*, 120–126.

Bibliography of this Dissertation's Publications

Publications of this dissertation

- D1 **Chen, Y.**; Kirchmair, J. Cheminformatics in Natural Product-Based Drug Discovery. *Mol. Inf.* **2020**, *39*, 2000171.
- D2 **Chen, Y.**; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* **2017**, *57* (9), 2099–2111.
- D3 **Chen, Y.**; de Bruyn Kops, C.; Kirchmair, J. Resources for Chemical, Biological, and Structural Data on Natural Products. In *Progress in the Chemistry of Organic Natural Products*; Kinghorn, A. D., Falk, H., Gibbons, S., Kobayashi, J., Asakawa, Y., Liu, J.-K., Eds.; Springer, 2019; Vol. 110, pp 37–71.
- D4 **Chen, Y.**; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58* (8), 1518–1532.
- D5 **Chen, Y.**; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **2019**, *9* (2), 43.
- D6 **Chen, Y.**; Mathai, N.; Kirchmair, J. Scope of 3D Shape-Based Approaches in Predicting the Macromolecular Targets of Structurally Complex Small Molecules Including Natural Products and Macrocyclic Ligands. *J. Chem. Inf. Model.* **2020**, *60* (6), 2858–2875.

Additional publications

- A1 Stork, C.; Embruch, G.; Šícho, M.; de Bruyn Kops, C.; **Chen, Y.**; Svozil, D.; Kirchmair, J. NERDD: A Web Portal Providing Access to in silico Tools for Drug Discovery. *Bioinformatics* **2020**, *36* (4), 1291–1292.
- A2 Stork, C.; **Chen, Y.**; Šícho, M.; Kirchmair, J. Hit Dexter 2.0: Machine-Learning Models for the Prediction of Frequent Hitters. *J. Chem. Inf. Model.* **2019**, *59* (3), 1030–1043.
- A3 Langeder, J.; Grienke, U.; **Chen, Y.**; Kirchmair, J.; Schmidtke, M.; Rollinger, J. M. Natural Products against Acute Respiratory Infections: Strategies and Lessons Learned. *J. Ethnopharmacol.* **2020**, *248*, 112298.

Abbreviations

2D	two-dimensional
3D	three-dimensional
ADME	absorption, distribution, metabolism, and excretion
AUC	area under the receiver operating characteristic curve
CSM	complex small molecules
FN	false negatives
FP	false positives
MCC	Matthews correlation coefficient
NP	natural product
PC	principal component
PCA	principal component analysis
PDB	Protein Data Bank
SMARTS	SMILES arbitrary target specification
SMILES	Simplified Molecular Input Line Entry Specification
TN	true negatives
TP	true positives

Appendix A

This section is the supporting information for the publication:

Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58* (8), 1518–1532.

Supporting Information

Characterization of the Chemical Space of Known and Readily Obtainable Natural Products

*Ya Chen, Marina Garcia de Lomana, Nils-Ole Friedrich, Johannes Kirchmair**

Center for Bioinformatics, Department of Computer Science, Faculty of Mathematics,
Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany

*corresponding author
kirchmair@zbh.uni-hamburg.de
+49 (0) 40 42838 7303

TABLE OF CONTENTS

Table S1.	Loadings of the first two components resulting from PCA analysis	2
Figure S1.	Distribution of the number of saturated rings and aliphatic rings among known and readily obtainable NPs, as well as approved drugs	3
Figure S2.	Violin and box plots of the fraction of rotatable bonds of virtual NP databases, physical NP libraries, and the Newman and Cragg data set and NPs of PDB	4
Figure S3.	Violin and box plots of the fraction of C _{sp3} atoms of virtual NP databases, physical NP libraries, and the Newman and Cragg data set and NPs of PDB	5
Figure S4.	Histograms of the number of rotatable bonds for all virtual NP databases	6
Figure S5.	Histograms of the number of rotatable bonds for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB	7
Figure S6.	Histograms of the number of chiral centers for all virtual NP databases	8
Figure S7.	Histograms of the number of chiral centers for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB	9
Figure S8.	Histograms of the number of rings for all virtual NP databases	10
Figure S9.	Histograms of the number of rings for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB	11
Figure S10.	Histograms of the number of aromatic rings for all virtual NP databases	12
Figure S11.	Histograms of the number of aromatic rings for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB	13
Figure S12.	Histograms of the number of nitrogen atoms for all virtual NP databases	14
Figure S13.	Histograms of the number of nitrogen atoms for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB	15
Figure S14.	Histograms of the number of oxygen atoms for all virtual NP databases	16
Figure S15.	Histograms of the number of oxygen atoms for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB	17
Figure S16.	Histograms of the number of hydrogen-bond acceptors for all virtual NP databases	18
Figure S17.	Histograms of the number of the number of hydrogen-bond acceptors for all physical NP databases, the Newman and Cragg data set, and NPs of the PDB	19
Figure S18.	Histograms of the number of hydrogen-bond donors for all virtual NP databases	20
Figure S19.	Histograms of the number of hydrogen-bond donors for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB	21
Figure S20.	Histograms of the number of acidic atoms for all virtual NP databases	22
Figure S21.	Histograms of the number of acidic atoms for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB	23
Figure S22.	Histograms of the number of basic atoms for all virtual NP databases	24
Figure S23.	Histograms of the number of basic atoms for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB	25

Table S1. Loadings of the First Two Components Resulting from PCA Analysis.

Principal component	PC1	PC2
Eigenvalue	6.63	2.36
Variance (%)	39	14
Cumulative variance (%)	39	53
MW	0.36	0.11
a_heavy	0.36	0.12
a_nN	0.22	-0.03
a_nO	0.32	-0.21
Halogens	-0.02	-0.01
log <i>P</i> (o/w)	0.09	0.34
TPSA	0.35	-0.22
a_acc	0.35	-0.21
a_don	0.31	-0.26
b_rotR	-0.05	-0.12
a_acid	0.08	-0.40
a_base	0.19	0.23
FCharge	0.09	0.48
a_aro	0.22	0.00
chiral	0.24	0.24
FractionC _{sp3}	-0.01	0.22
NumRings	0.28	0.29

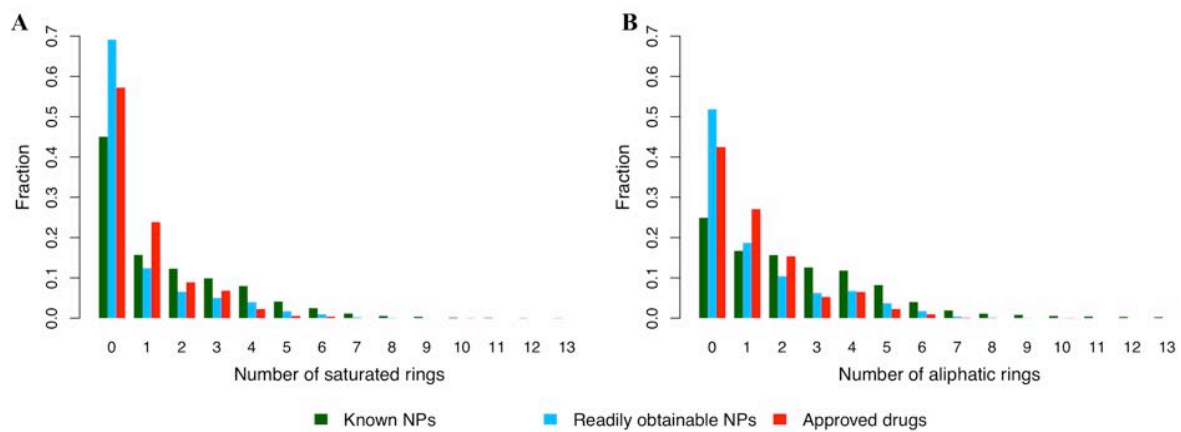


Figure S1. Distribution of the number of (A) saturated rings and (B) aliphatic rings among known and readily obtainable NPs, as well as approved drugs.

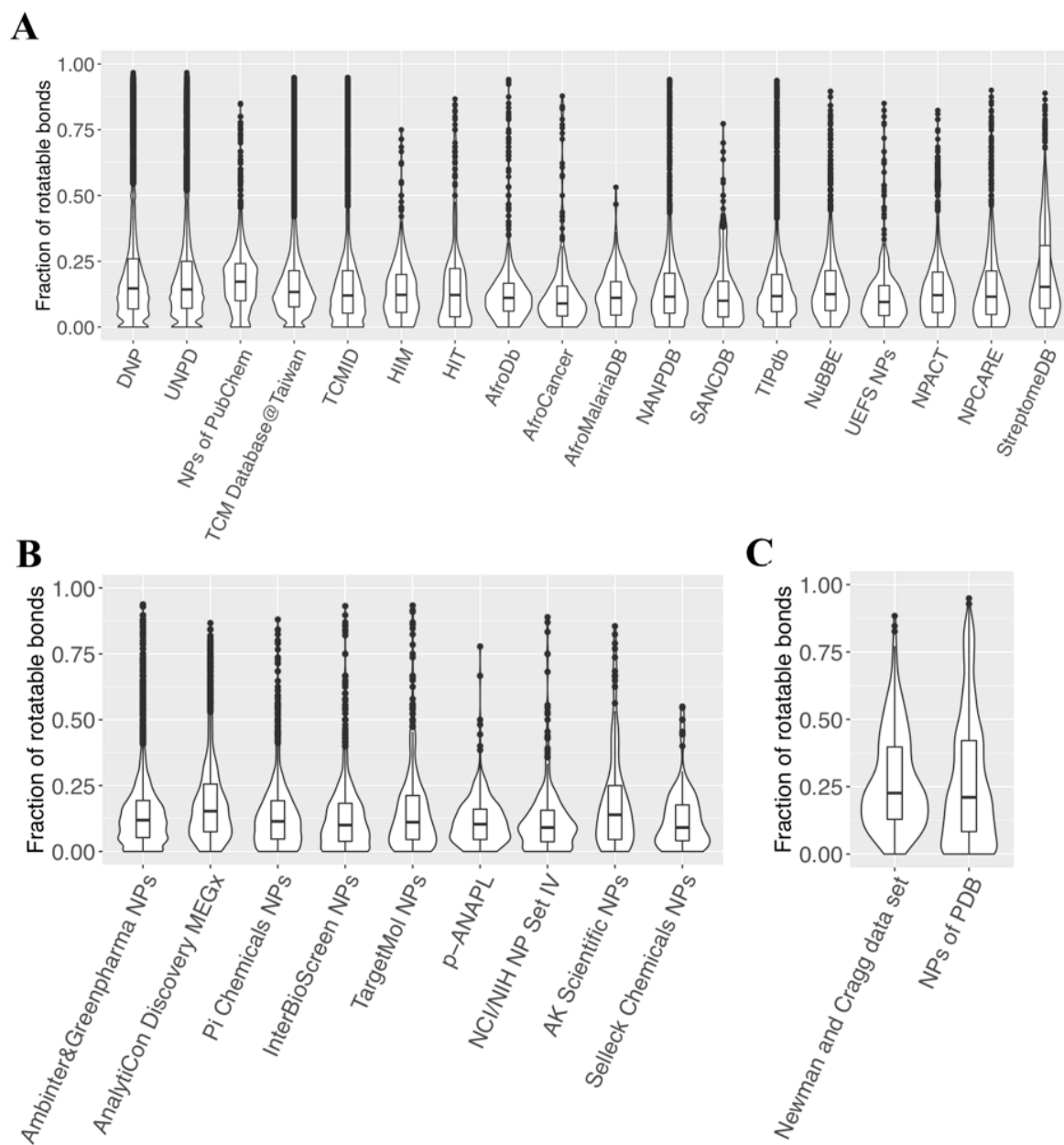


Figure S2. Violin and box plots of the fraction of rotatable bonds of (A) virtual NP databases, (B) physical NP libraries, and (C) the Newman and Cragg data set and NPs of PDB.

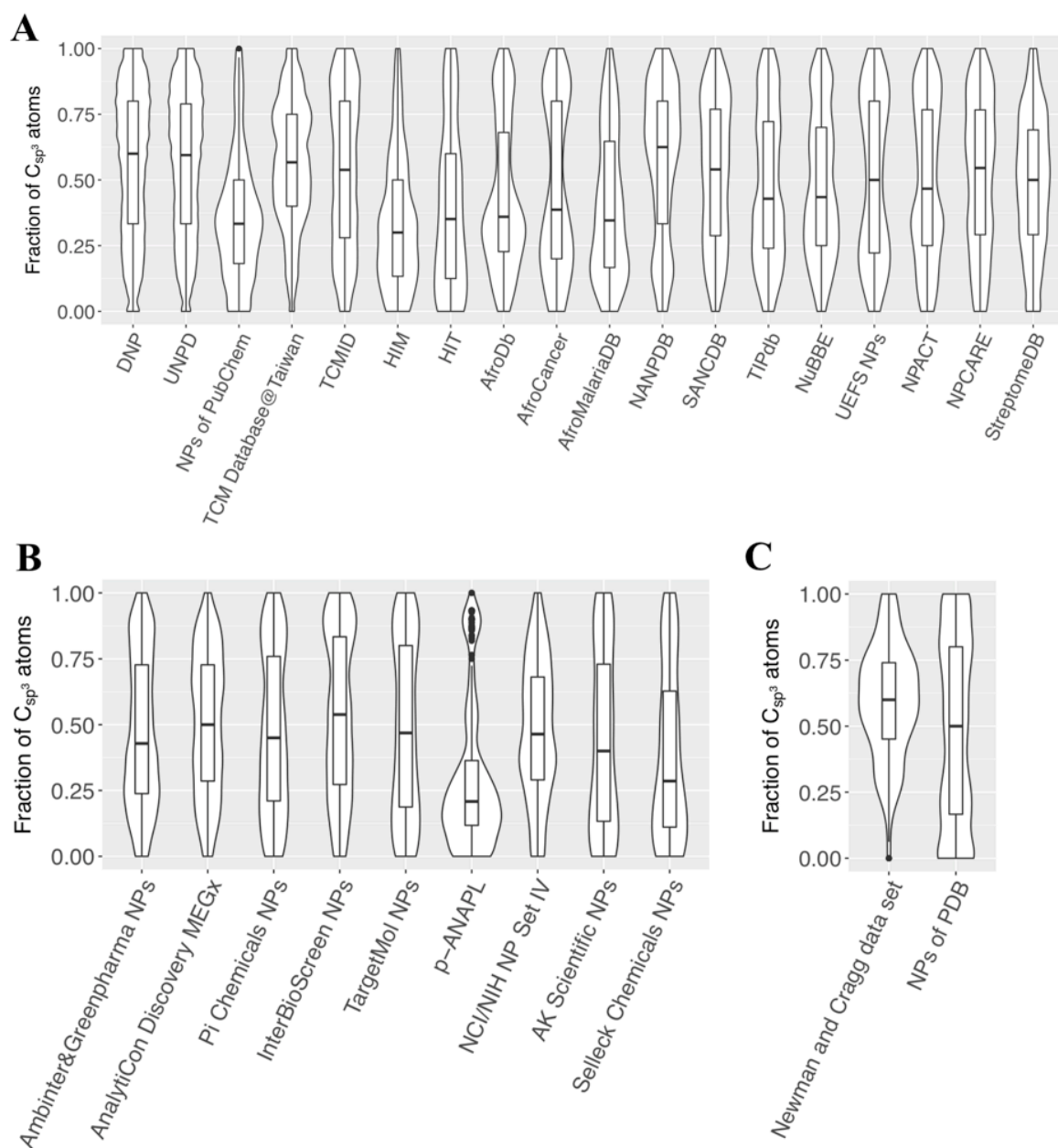


Figure S3. Violin and box plots of the fraction of C_{sp^3} atoms of (A) virtual NP databases, (B) physical NP libraries, and (C) the Newman and Cragg data set and NPs of PDB.

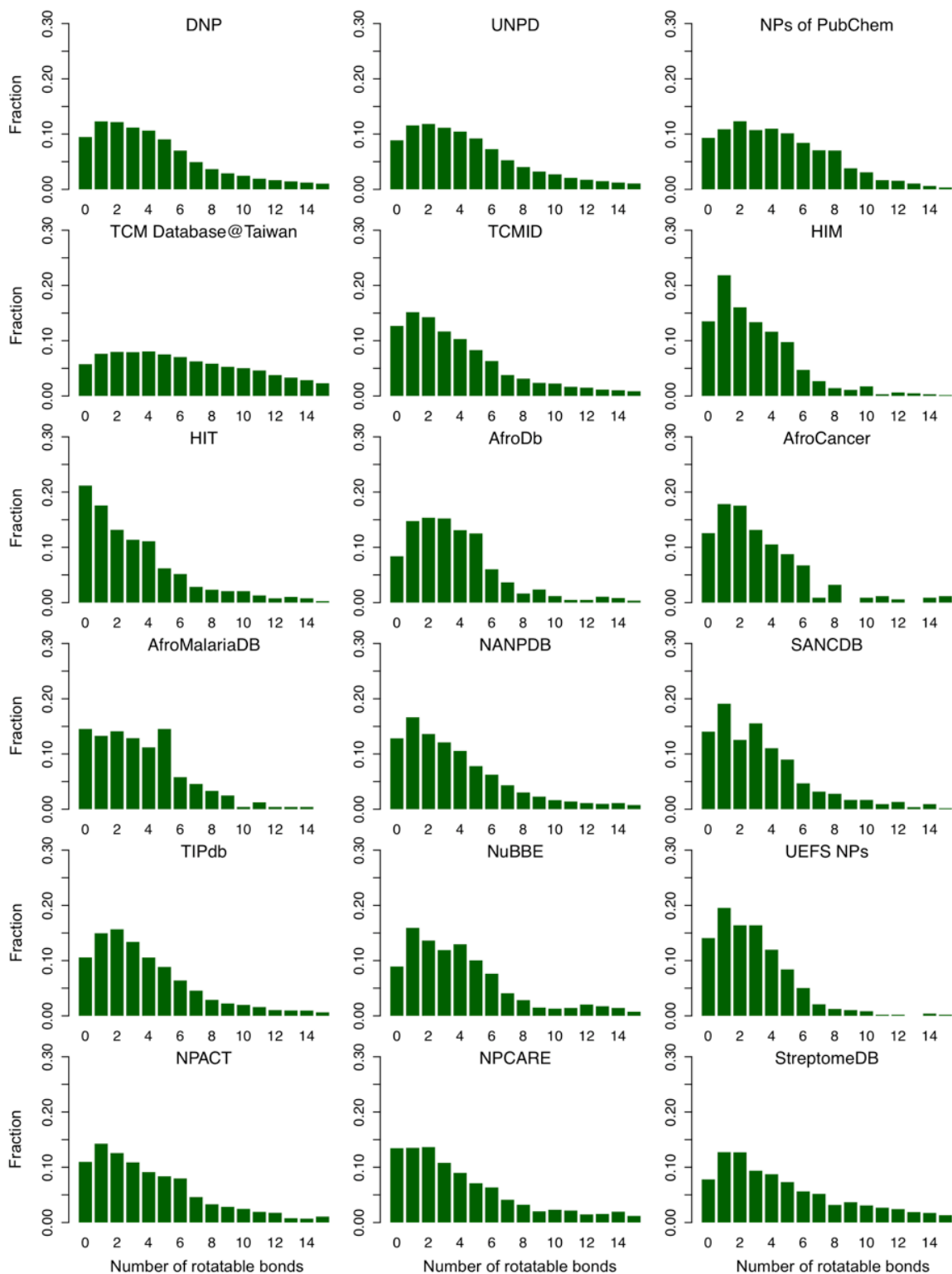


Figure S4. Histograms of the number of rotatable bonds for all virtual NP databases.

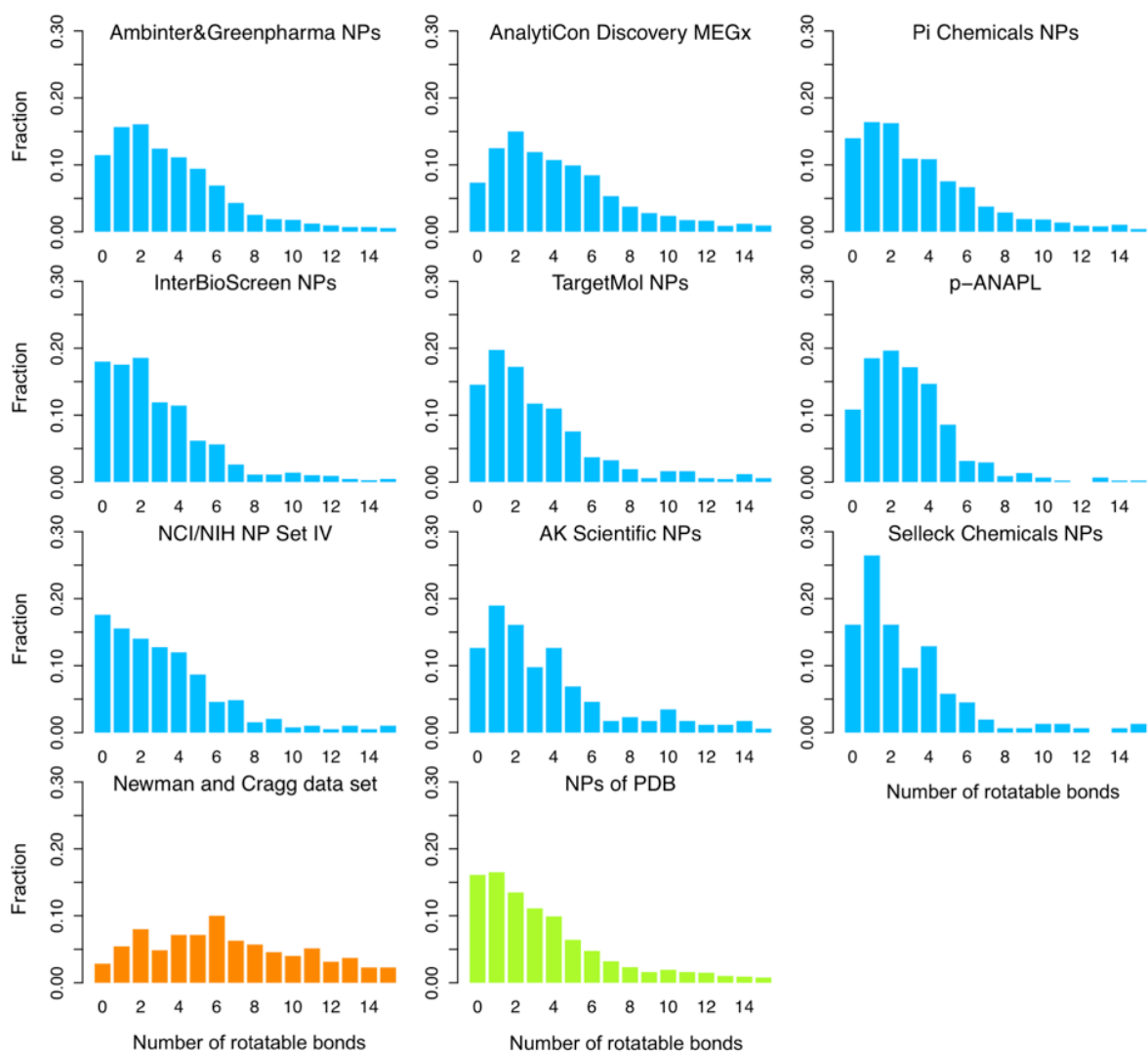


Figure S5. Histograms of the number of rotatable bonds for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB.

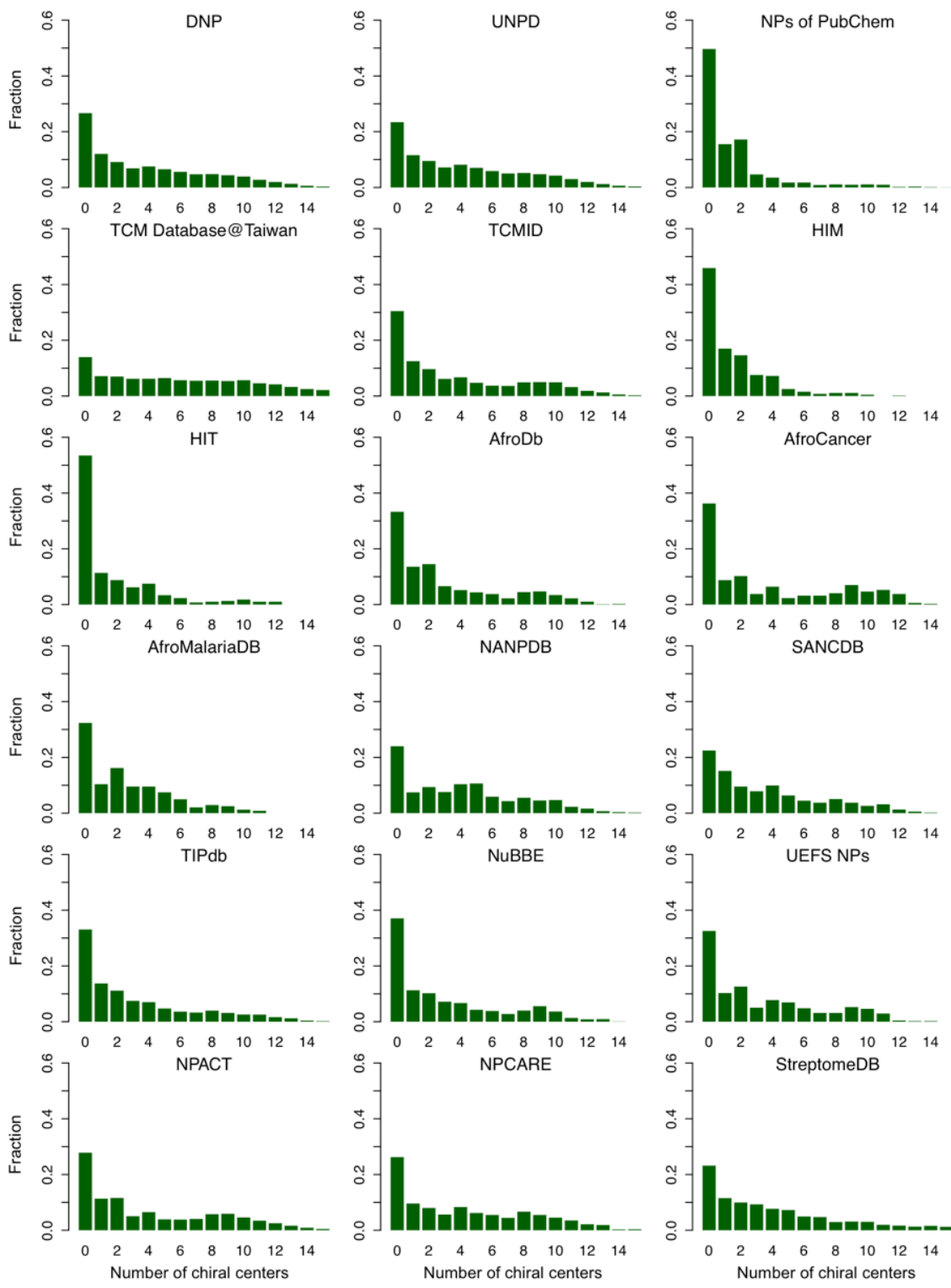


Figure S6. Histograms of the number of chiral centers for all virtual NP databases.

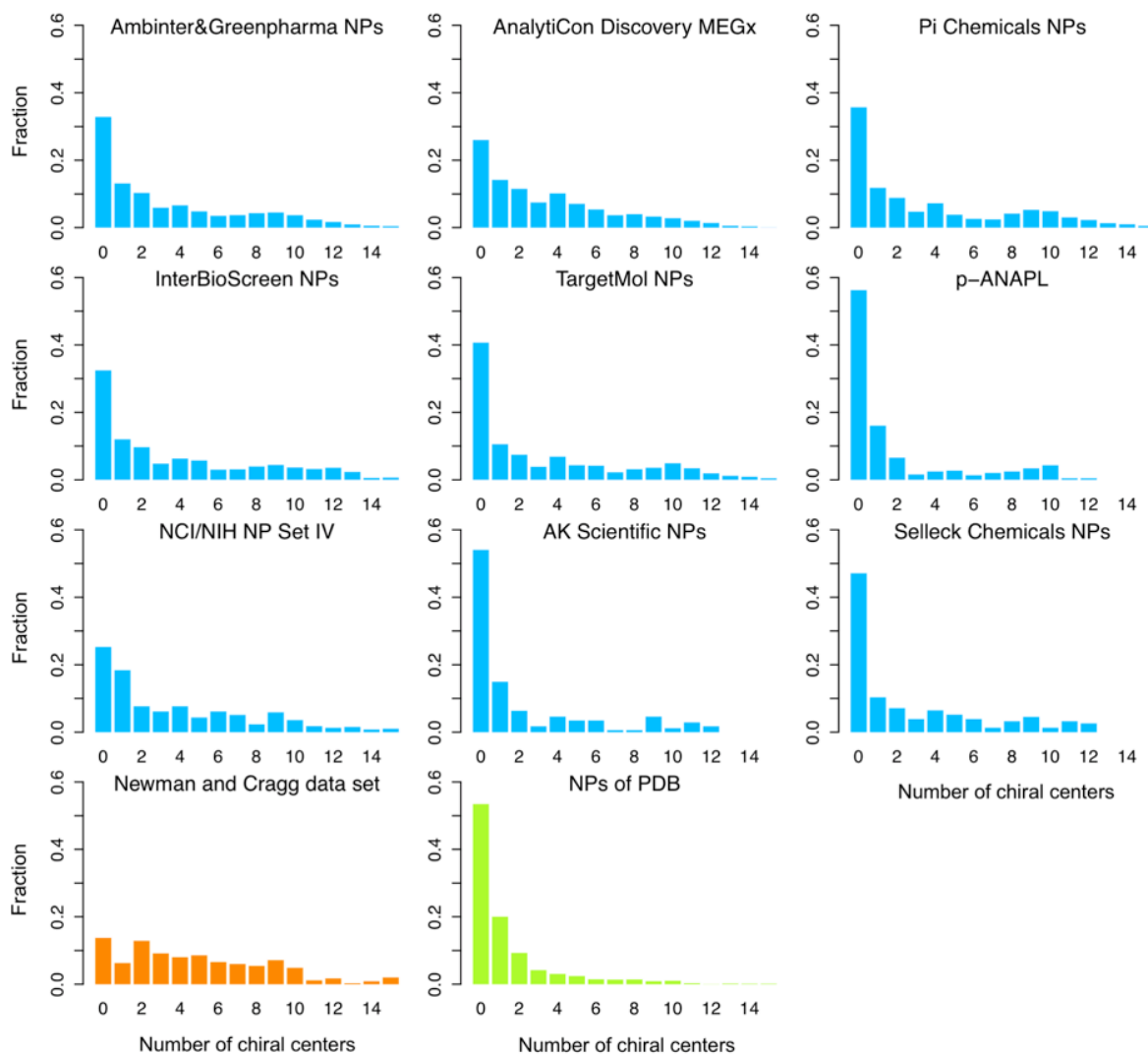


Figure S7. Histograms of the number of chiral centers for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB.

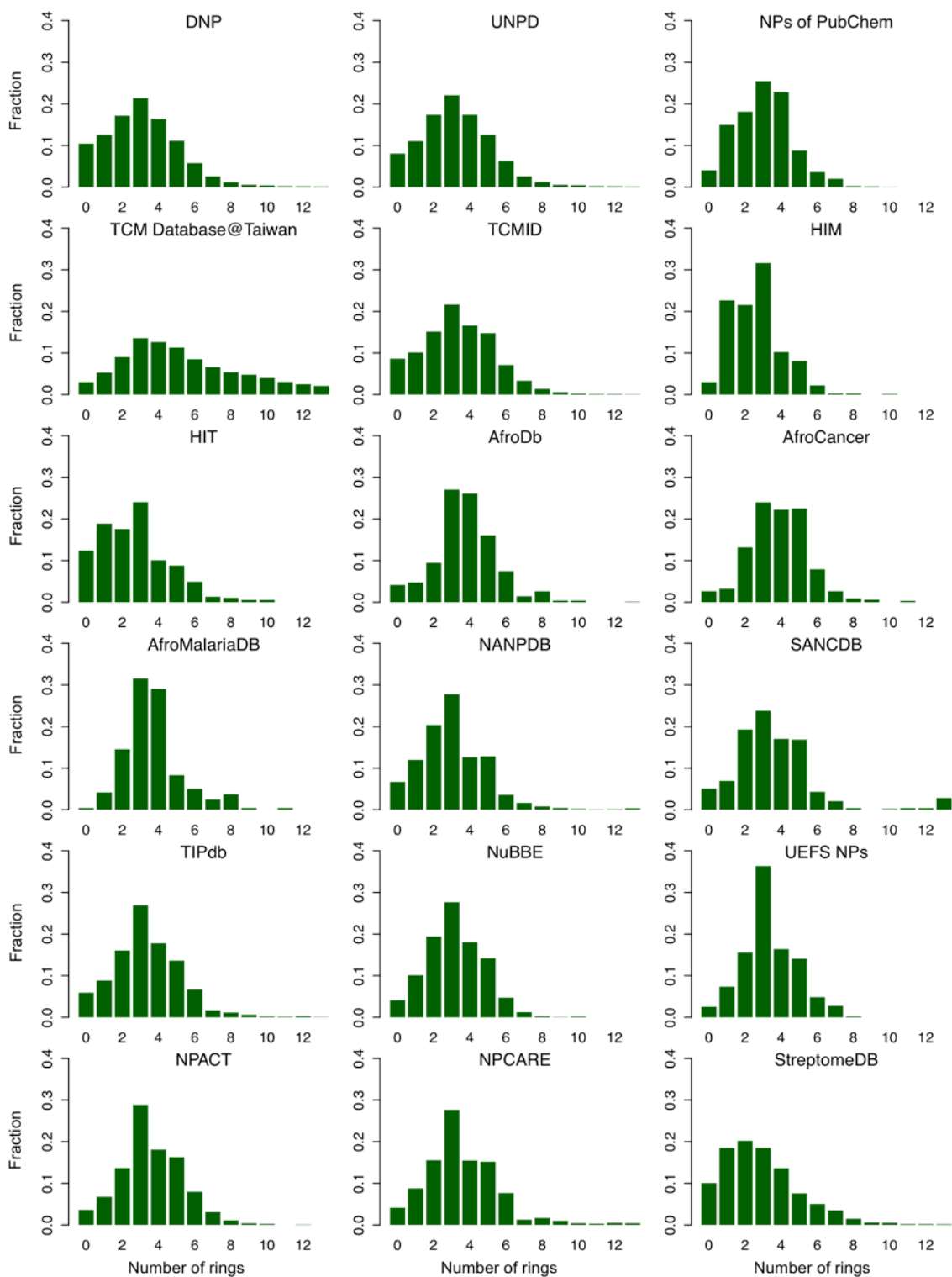


Figure S8. Histograms of the number of rings for all virtual NP databases.

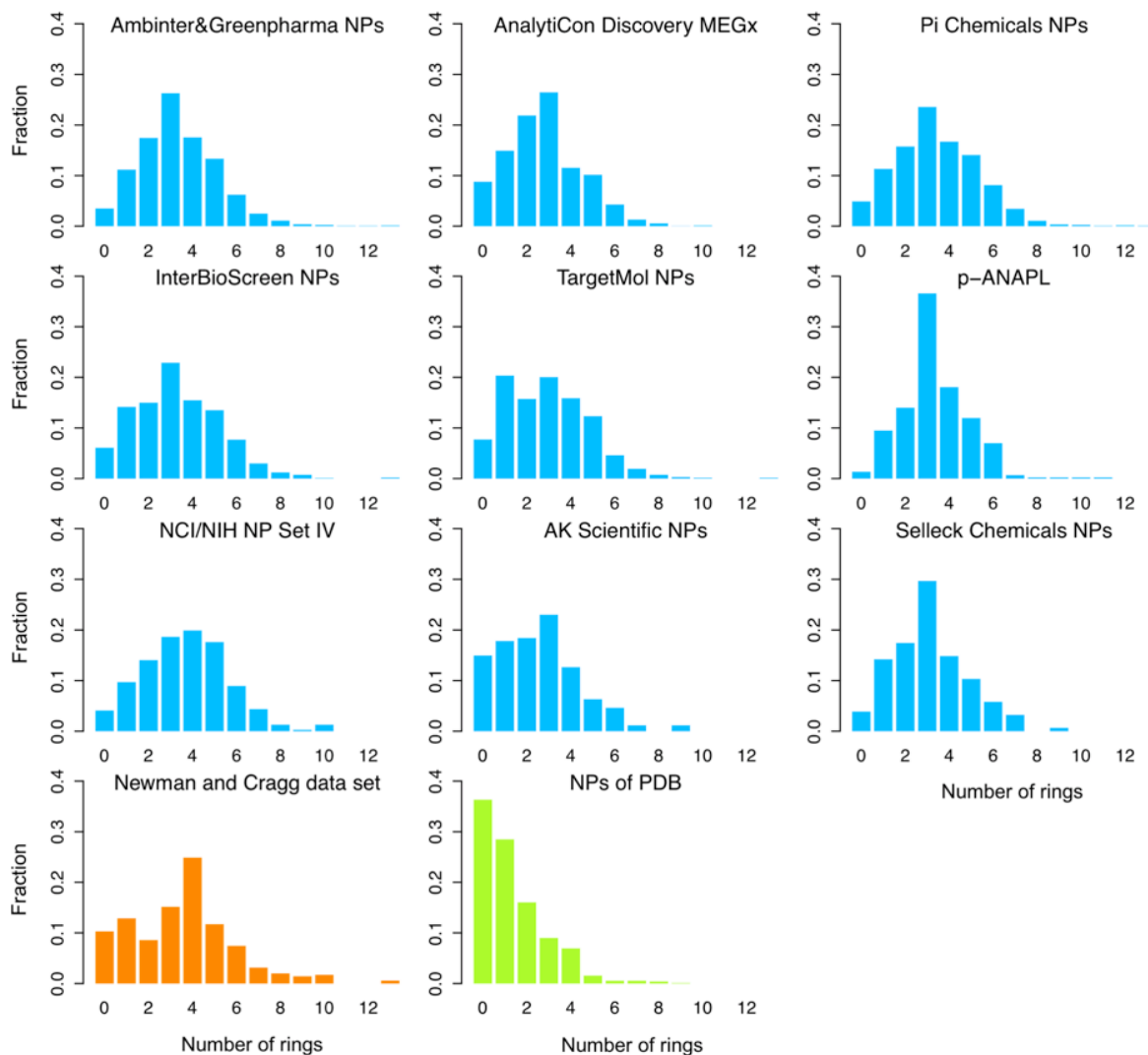


Figure S9. Histograms of the number of rings for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB.

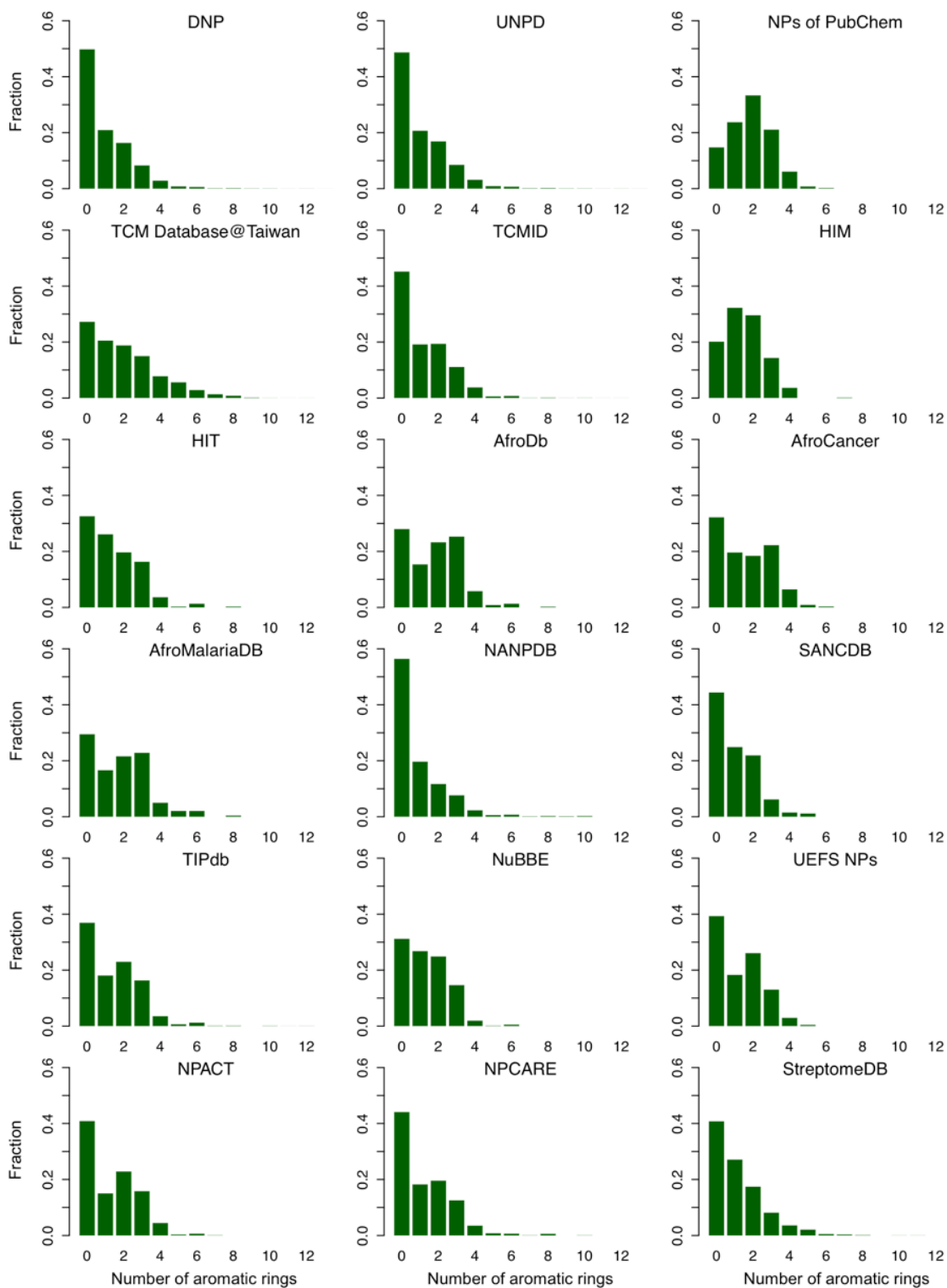


Figure S10. Histograms of the number of aromatic rings for all virtual NP databases.

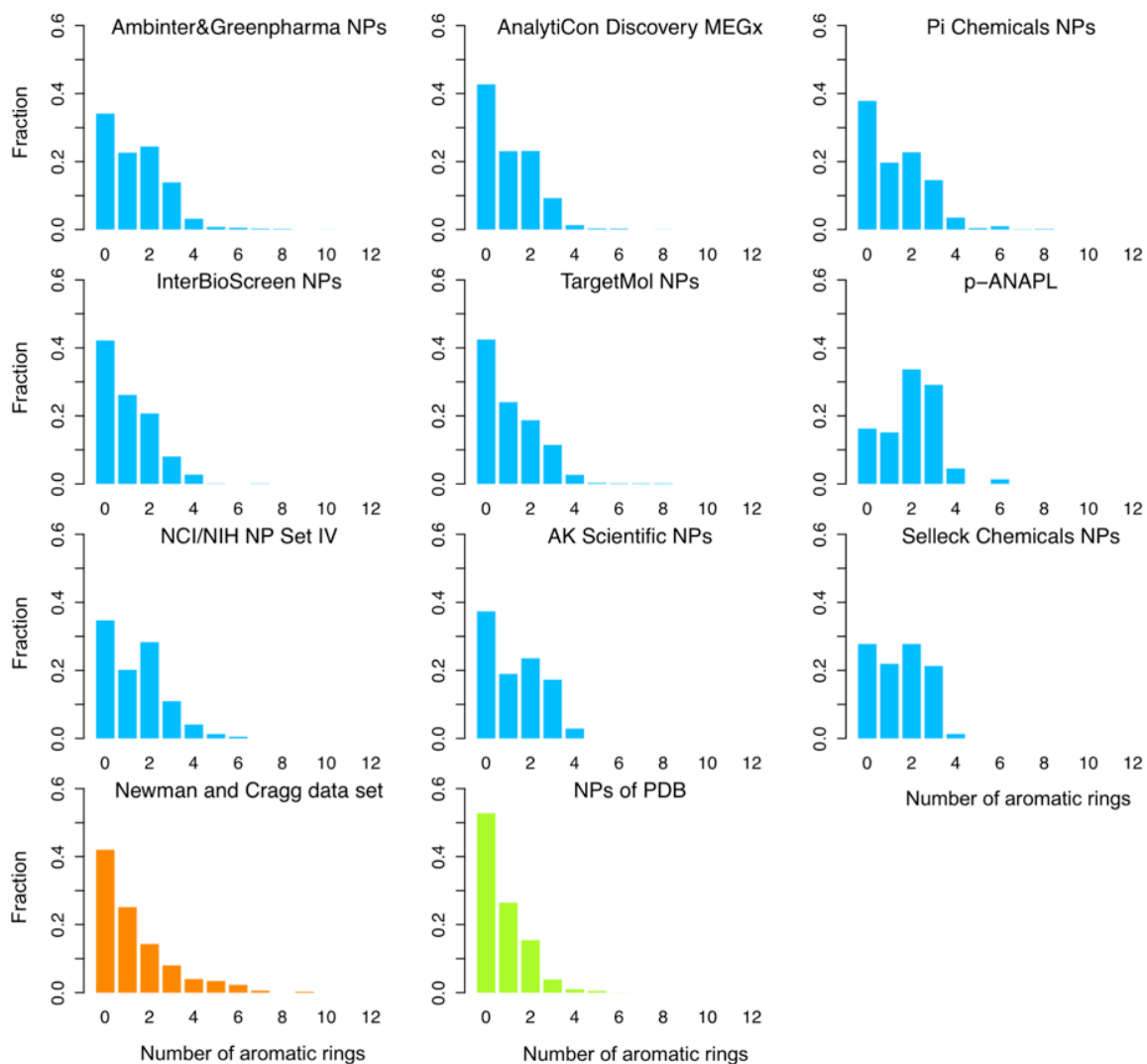


Figure S11. Histograms of the number of aromatic rings for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB.

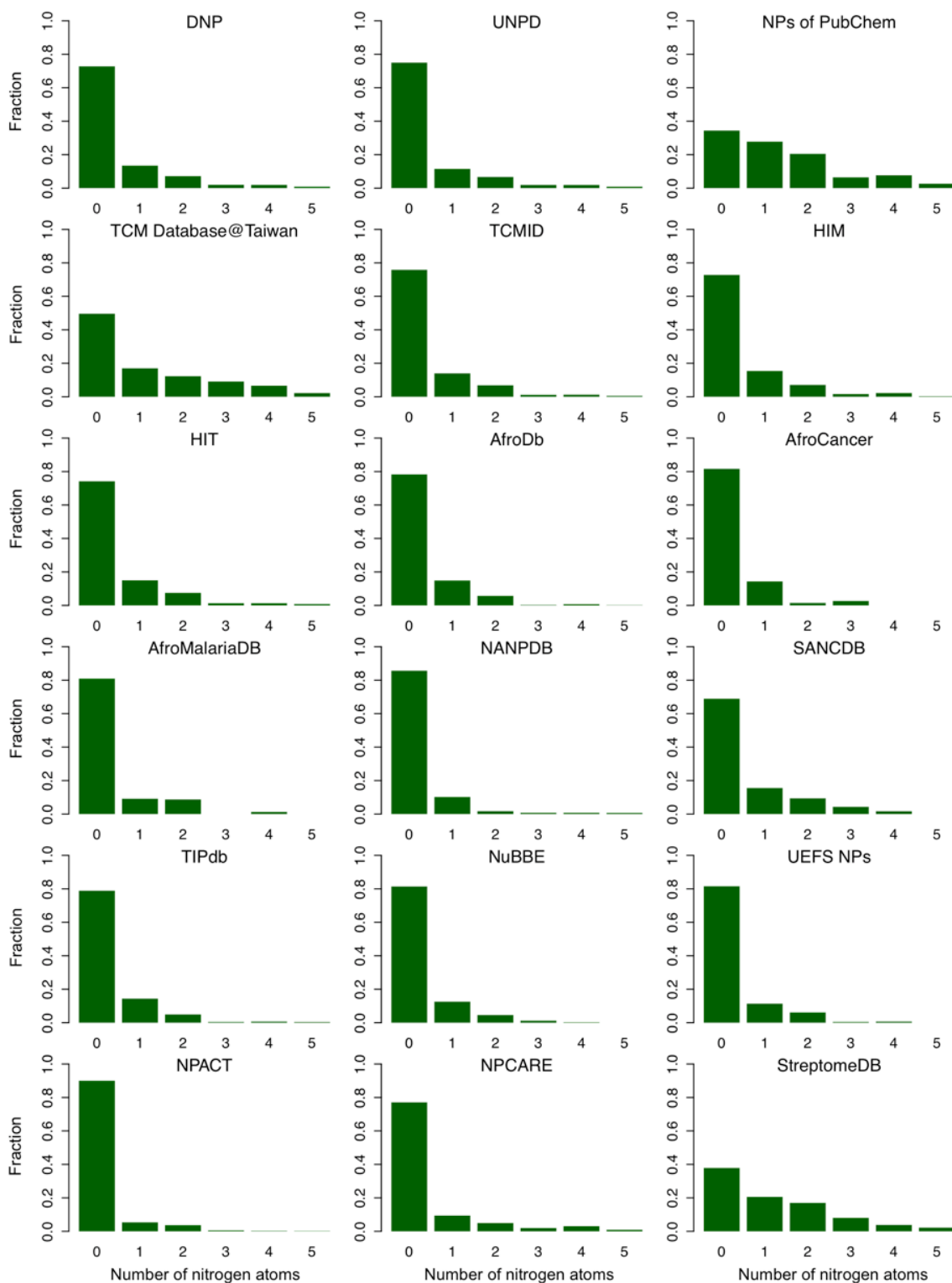


Figure S12. Histograms of the number of nitrogen atoms for all virtual NP databases.

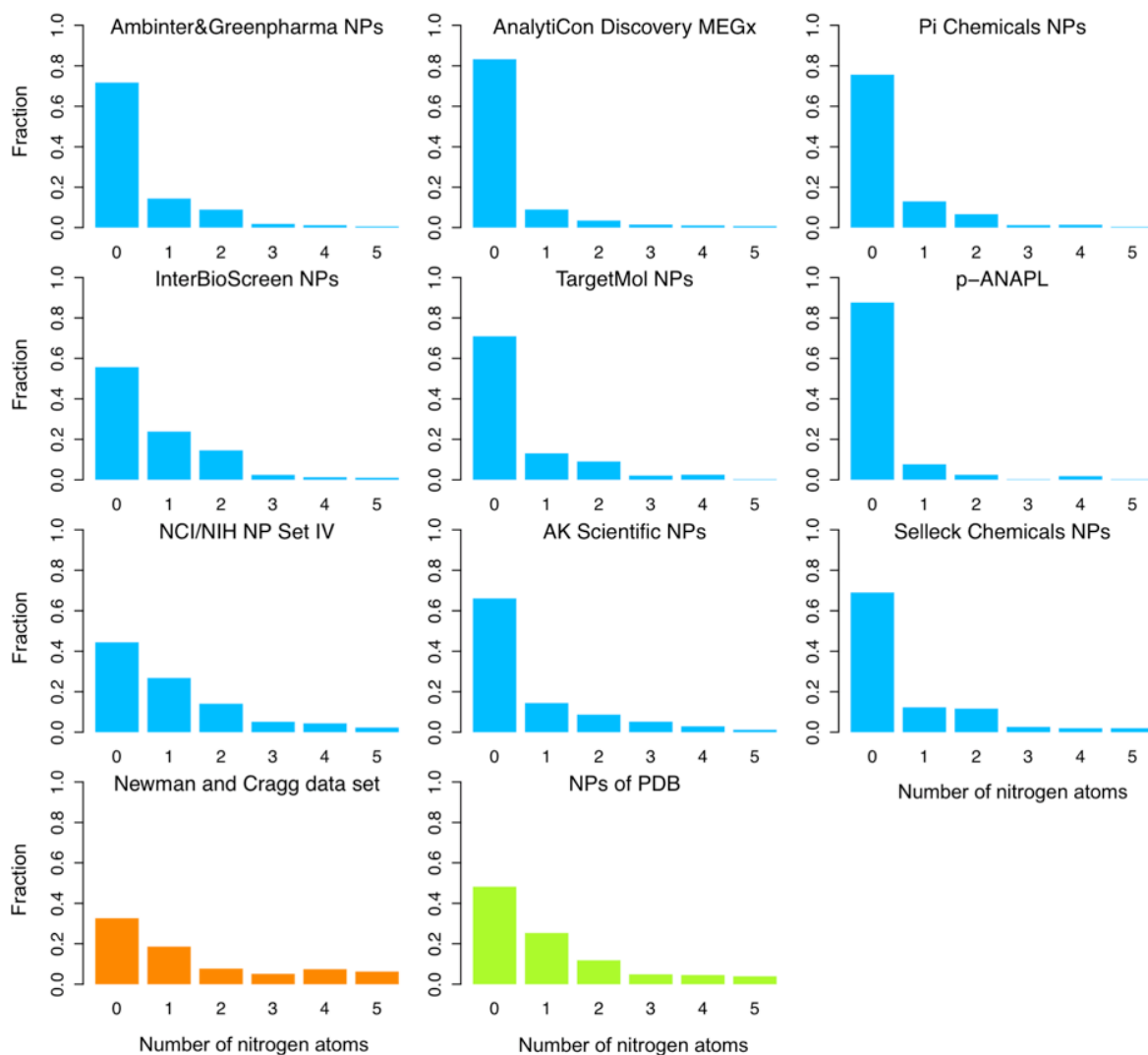


Figure S13. Histograms of the number of nitrogen atoms for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB.

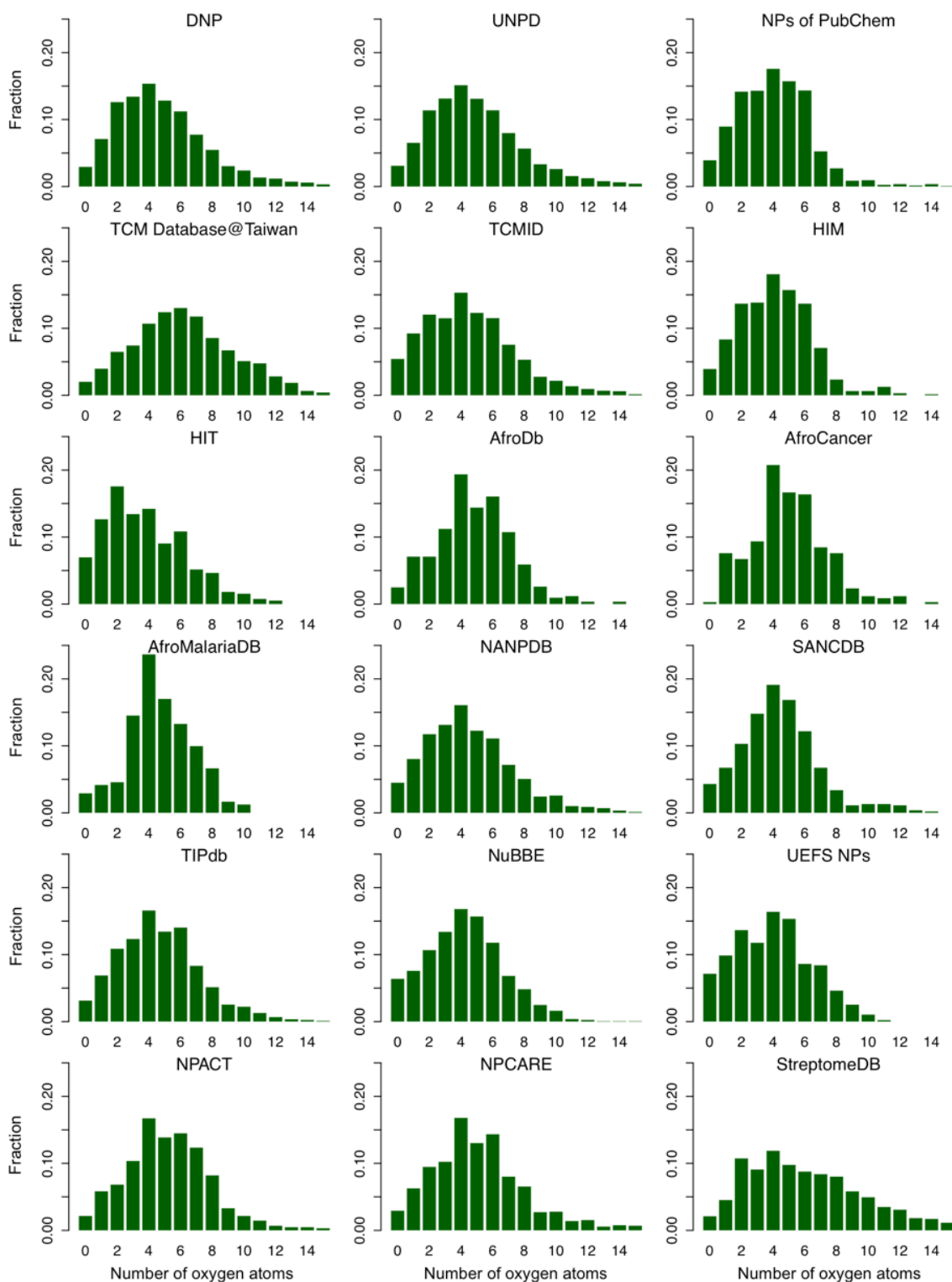


Figure S14. Histograms of the number of oxygen atoms for all virtual NP databases.

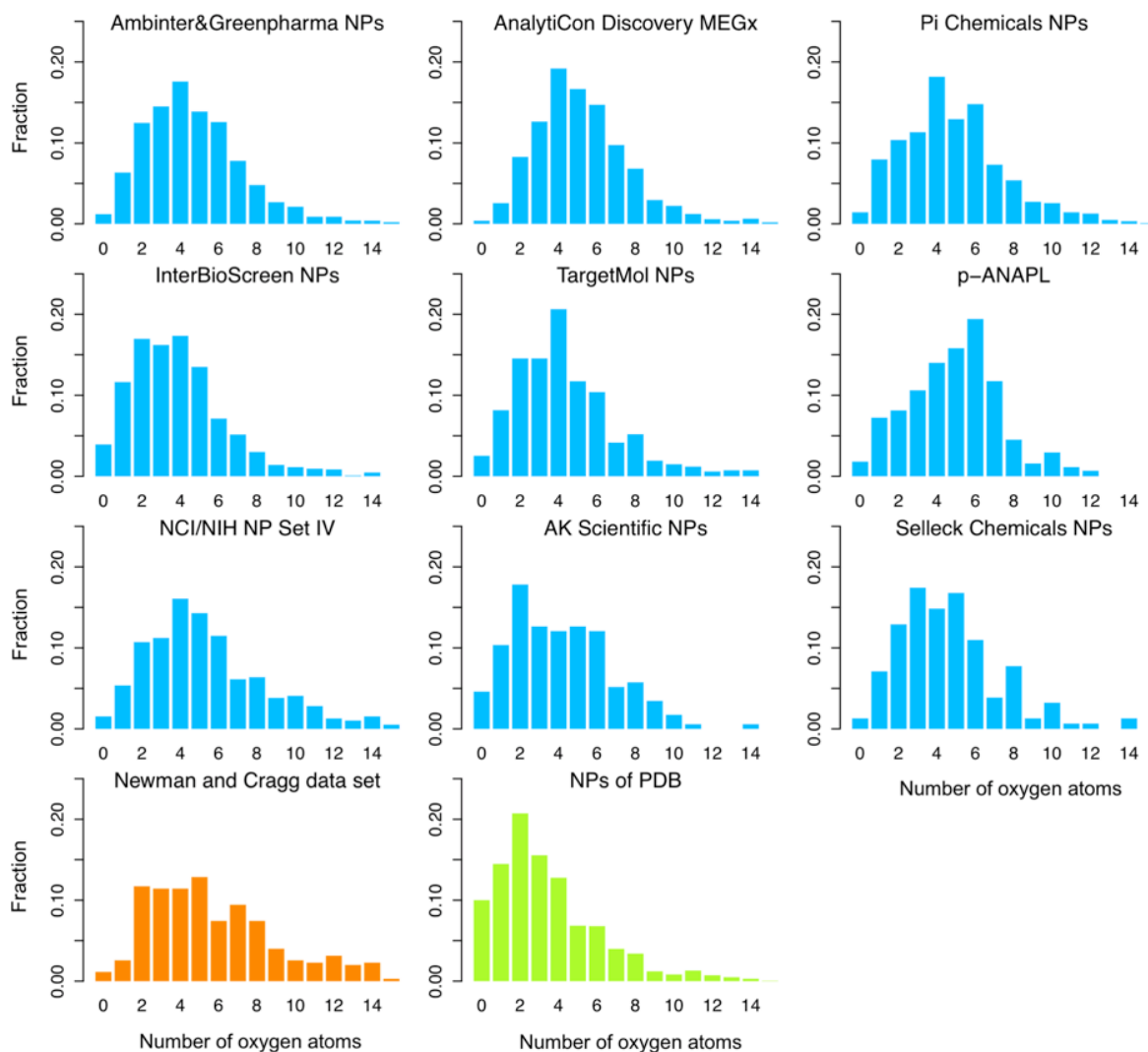


Figure S15. Histograms of the number of oxygen atoms for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB.

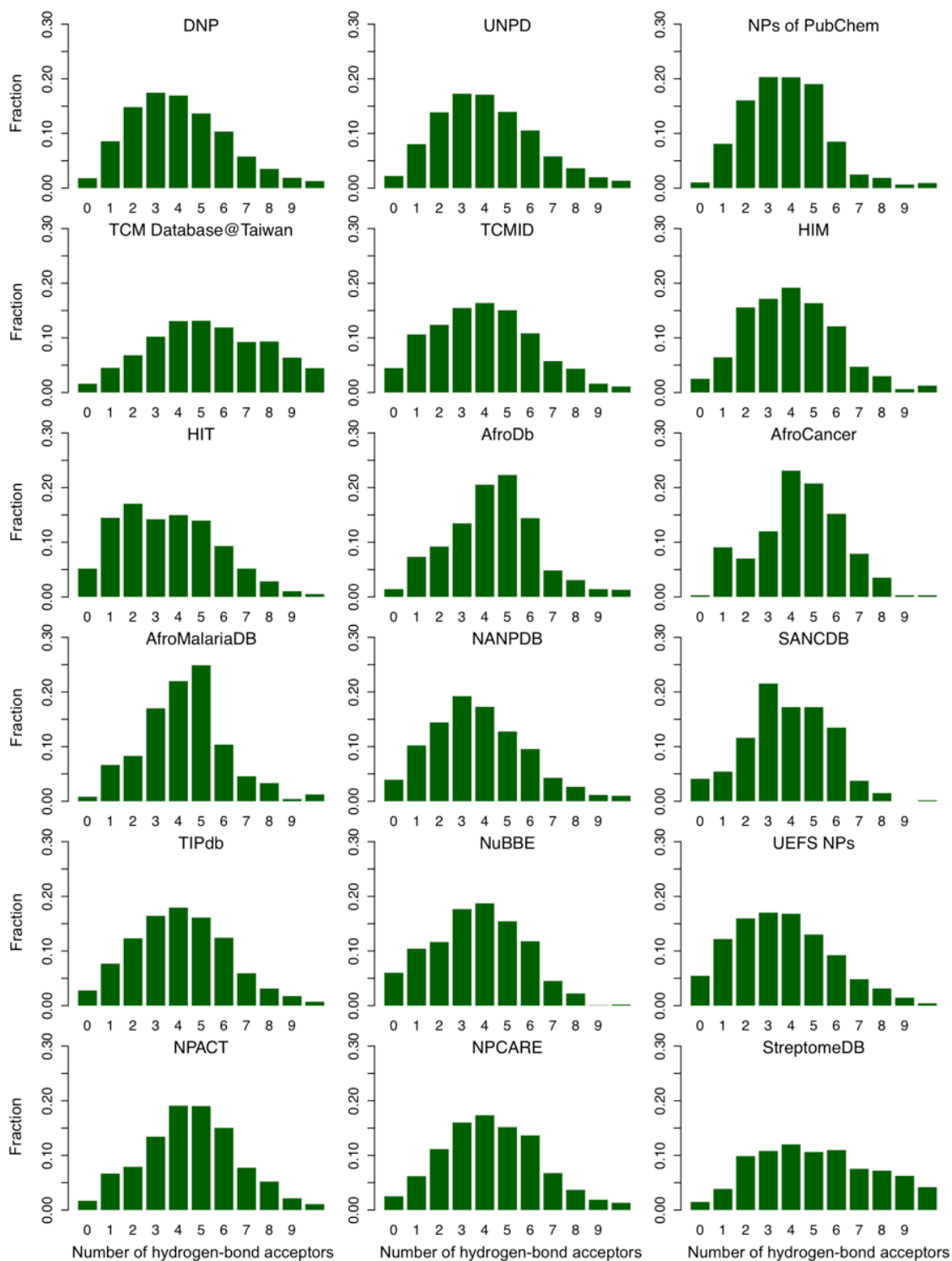


Figure S16. Histograms of the number of hydrogen-bond acceptors for all virtual NP databases.

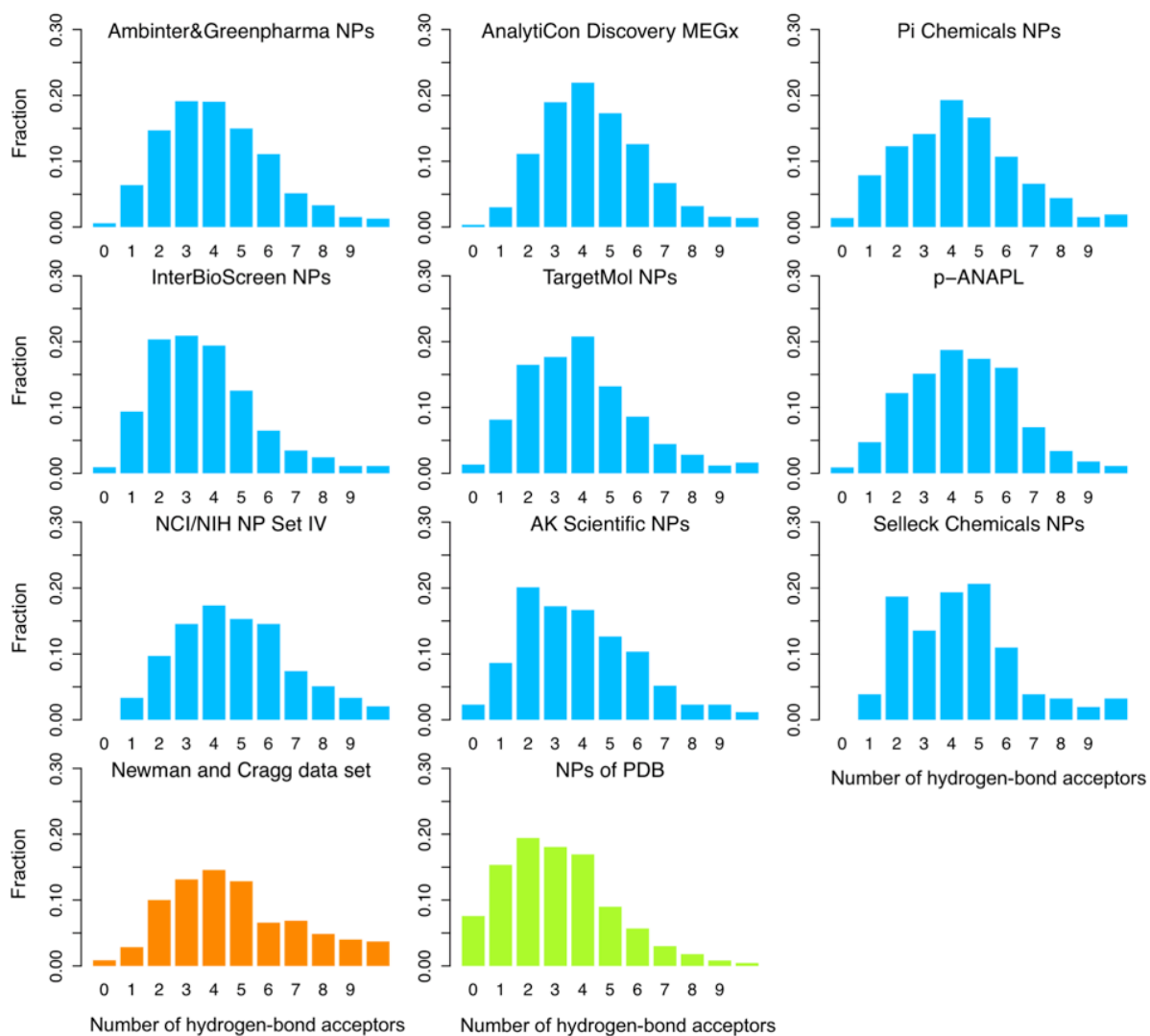


Figure S17. Histograms of the number of the number of hydrogen-bond acceptors for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB.

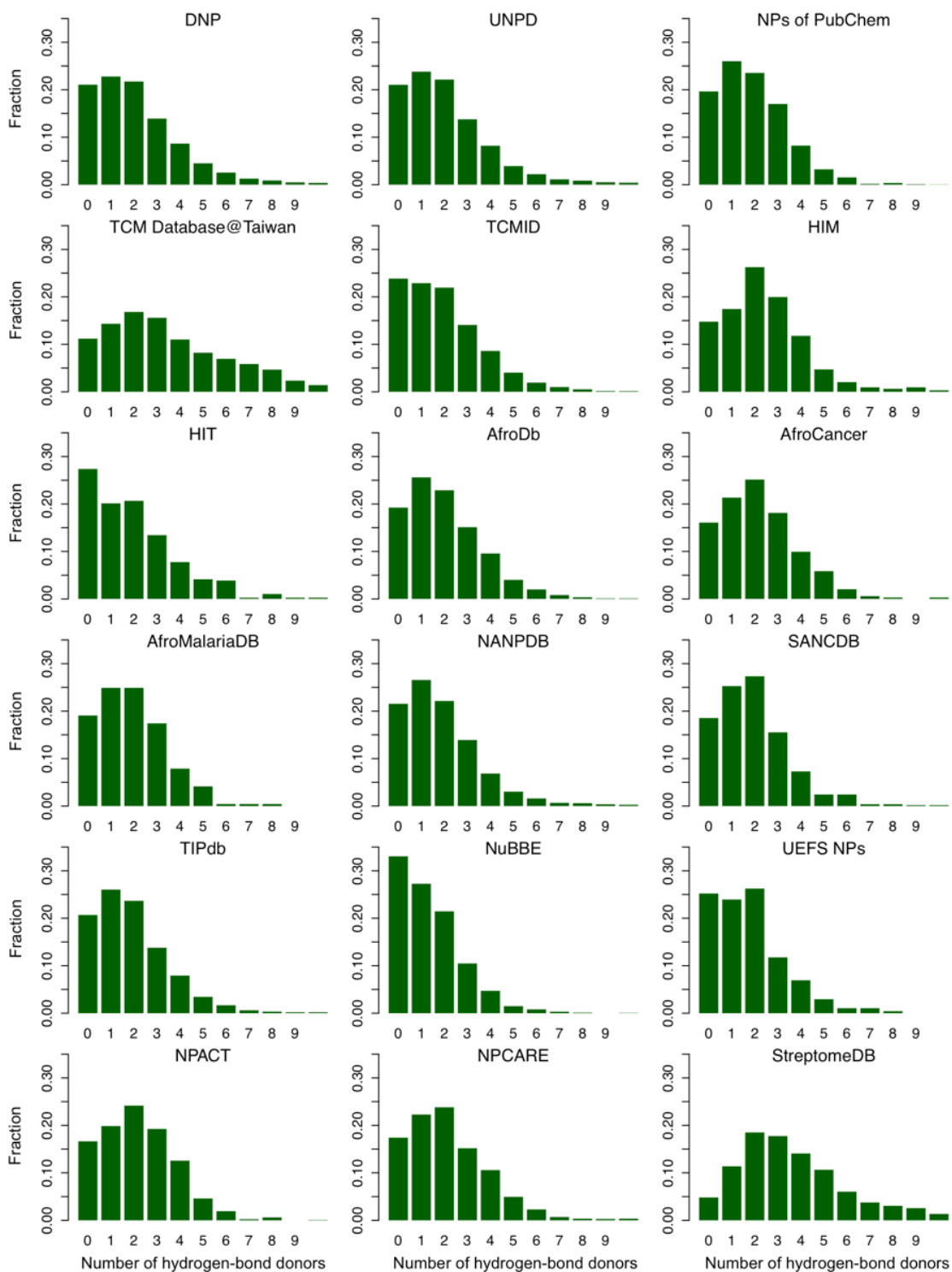


Figure S18. Histograms of the number of hydrogen-bond donors for all virtual NP databases.

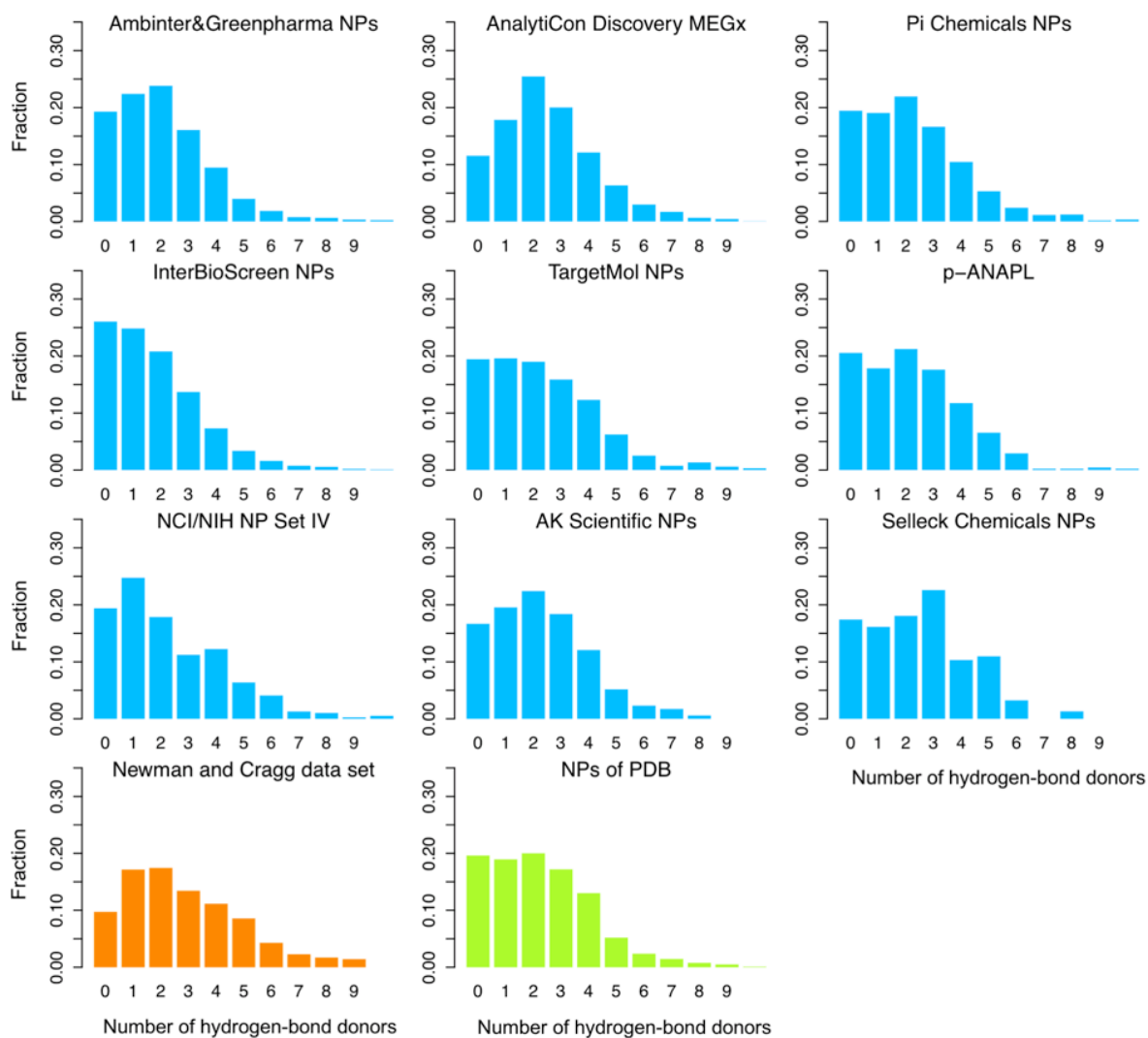


Figure S19. Histograms of the number of hydrogen-bond donors for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB.

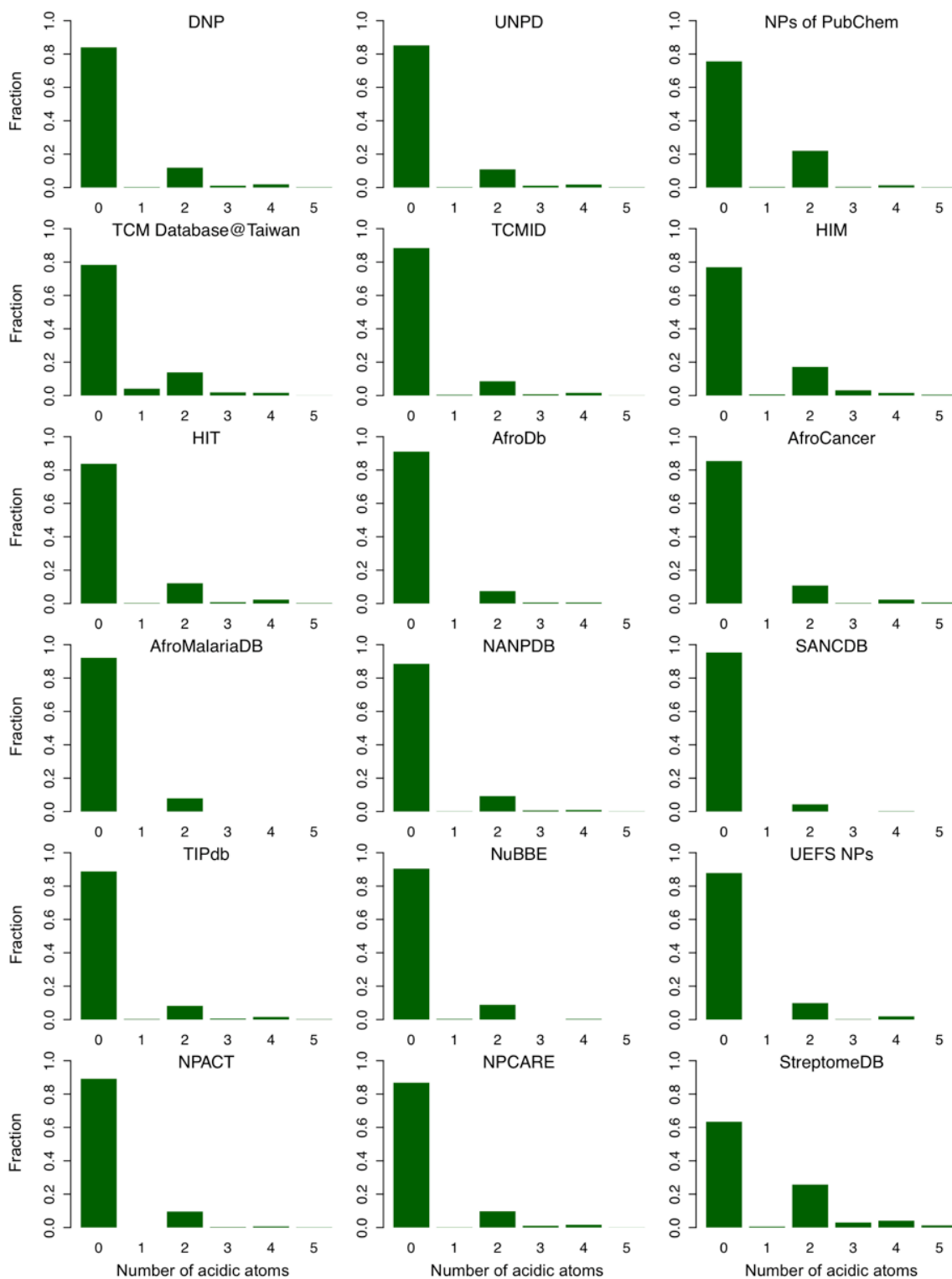


Figure S20. Histograms of the number of acidic atoms for all virtual NP databases.

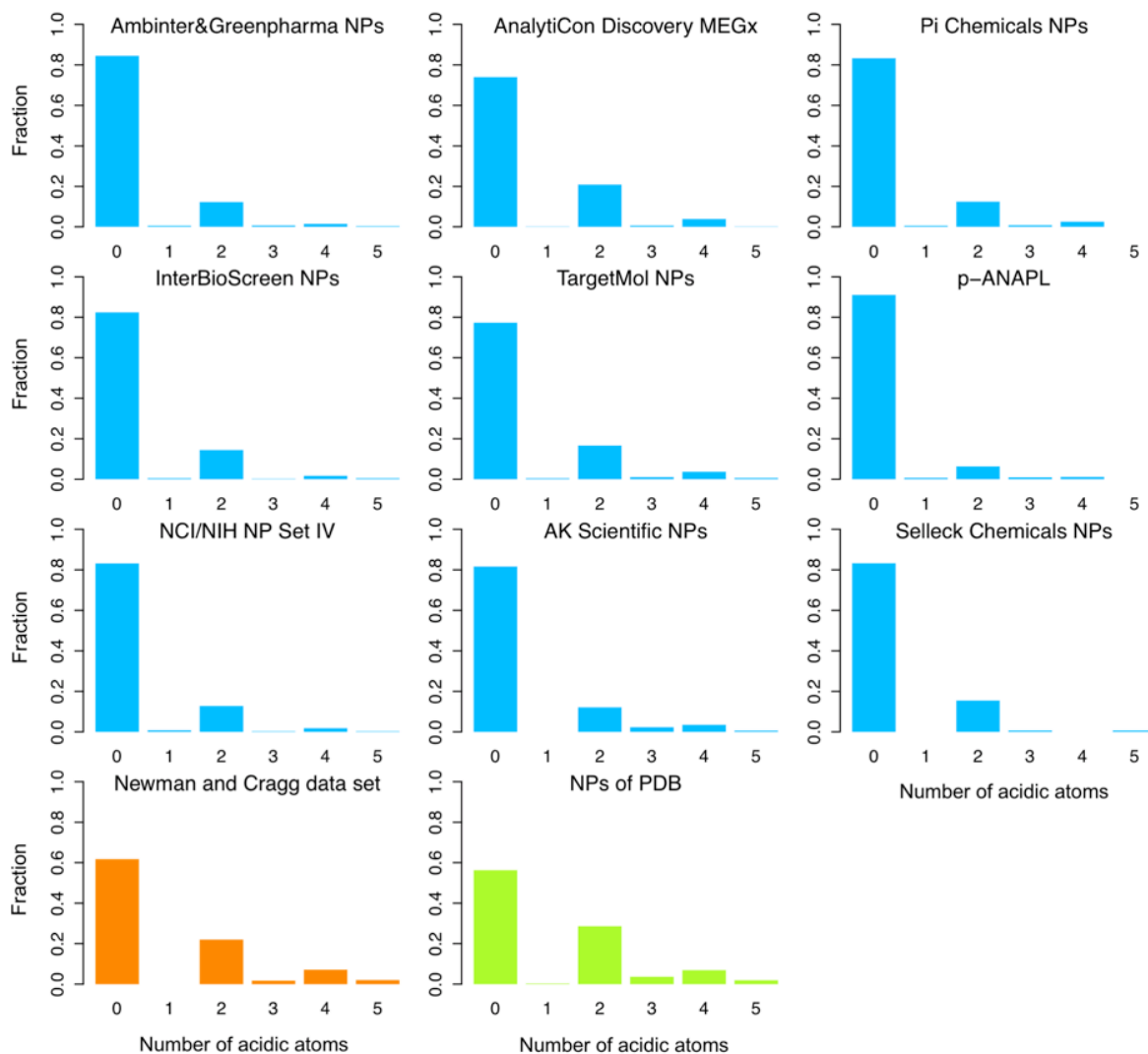


Figure S21. Histograms of the number of acidic atoms for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB.

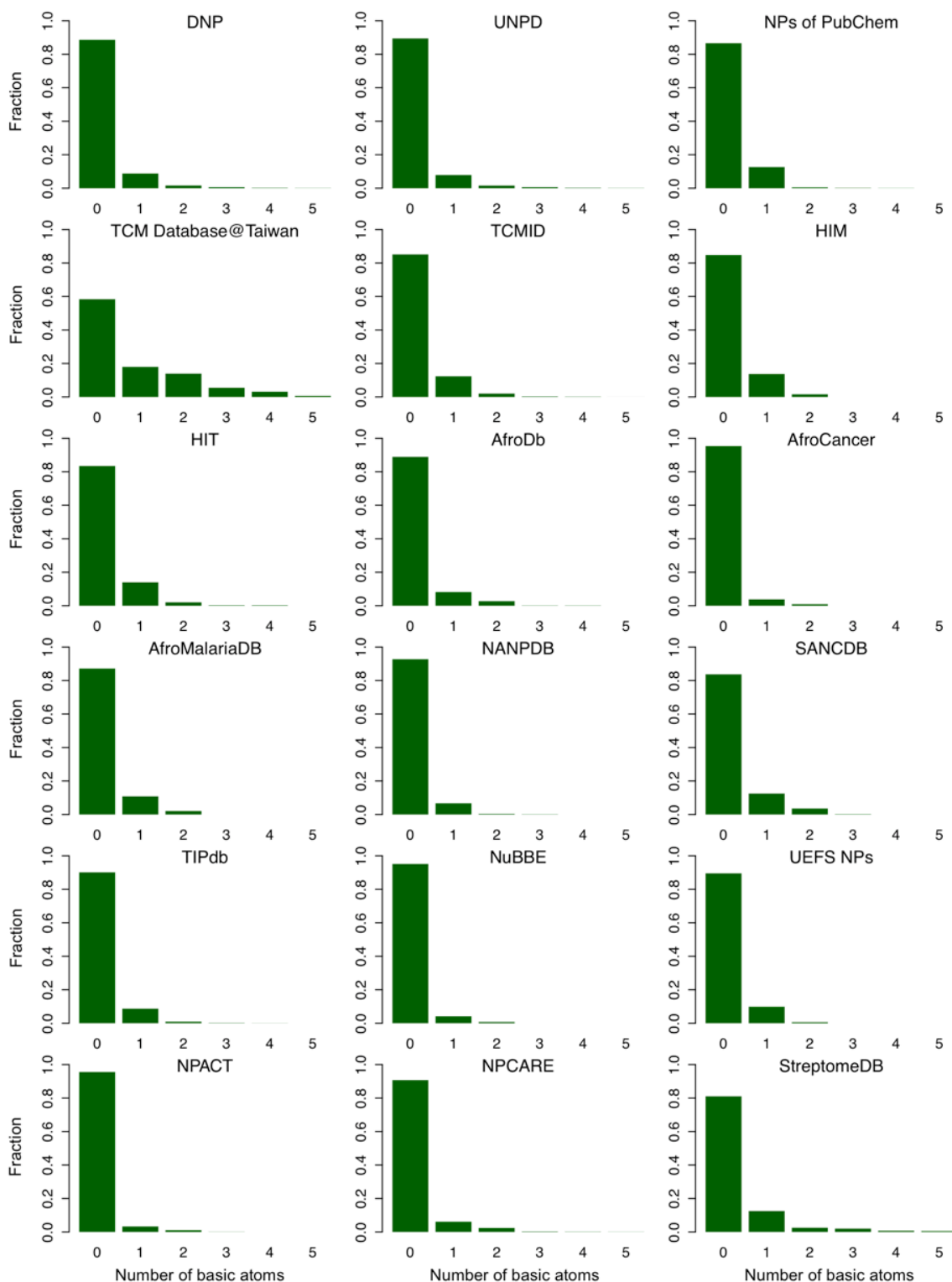


Figure S22. Histograms of the number of basic atoms for all virtual NP databases.

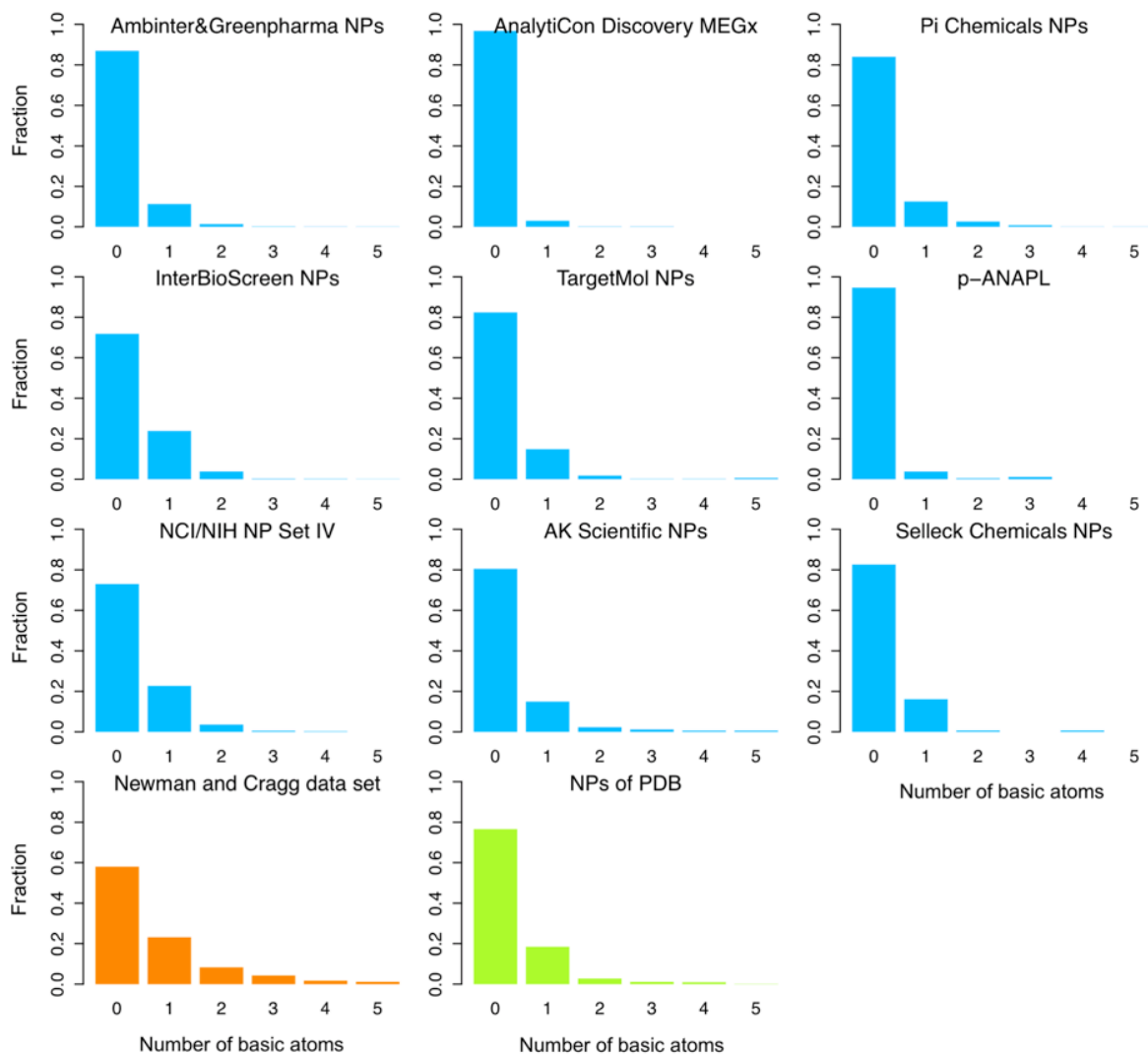


Figure S23. Histograms of the number of basic atoms for all physical NP libraries, the Newman and Cragg data set, and NPs of the PDB.

Appendix B

This section is the supporting information for the publication:

Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **2019**, *9* (2), 43.

Supporting Information for

NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules

Ya Chen ¹, Conrad Stork ¹, Steffen Hirte ¹ and Johannes Kirchmair ^{1,2,3,*}

¹ Center for Bioinformatics (ZBH), Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany; chen@zbh.uni-hamburg.de (Y.C.); stork@zbh.uni-hamburg.de (C.S.); steffen.hirte@studium.uni-hamburg.de (S.H.)

² Department of Chemistry, University of Bergen, 5007 Bergen, Norway

³ Computational Biology Unit (CBU), Department of Informatics, University of Bergen, 5008 Bergen, Norway

* Correspondence: johannes.kirchmair@uib.no or kirchmair@zbh.uni-hamburg.de; Tel.: +47-5558-3464

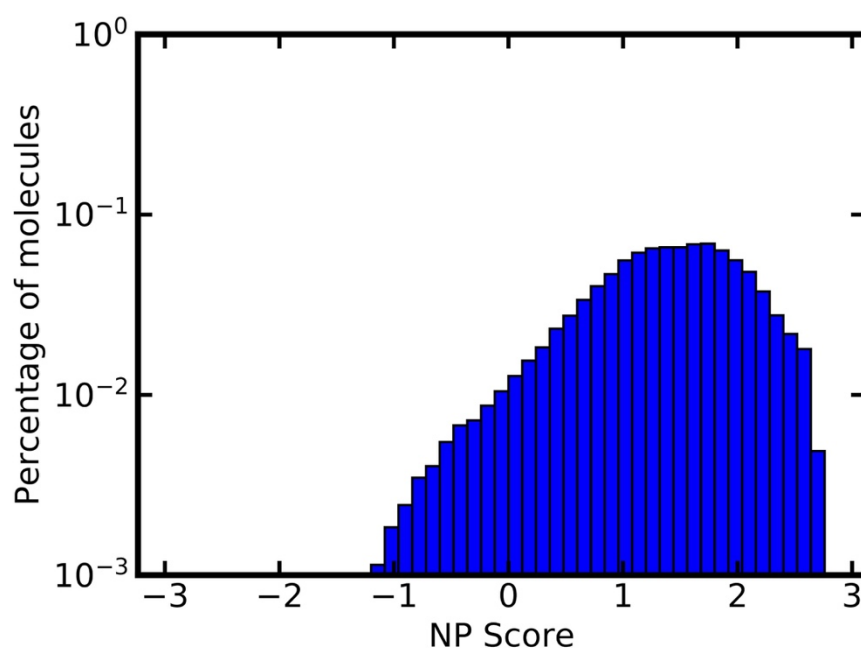


Figure S1. Distribution of calculated NP-likeness scores for the DNP (after removal of any compounds present in the training set). Note that the y-axis is in logarithmic scale.

Appendix C

This section is the supporting information for the publication:

Chen, Y.; Mathai, N.; Kirchmair, J. Scope of 3D Shape-Based Approaches in Predicting the Macromolecular Targets of Structurally Complex Small Molecules Including Natural Products and Macrocyclic Ligands. *J. Chem. Inf. Model.* **2020**, *60* (6), 2858-2875.

Supporting Information

Scope of 3D shape-based approaches in predicting the macromolecular targets of structurally complex small molecules including natural products and macrocyclic ligands

Ya Chen,¹ Neann Mathai² and Johannes Kirchmair^{1,2,3}*

¹ Center for Bioinformatics (ZBH), Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany

² Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway

³ Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

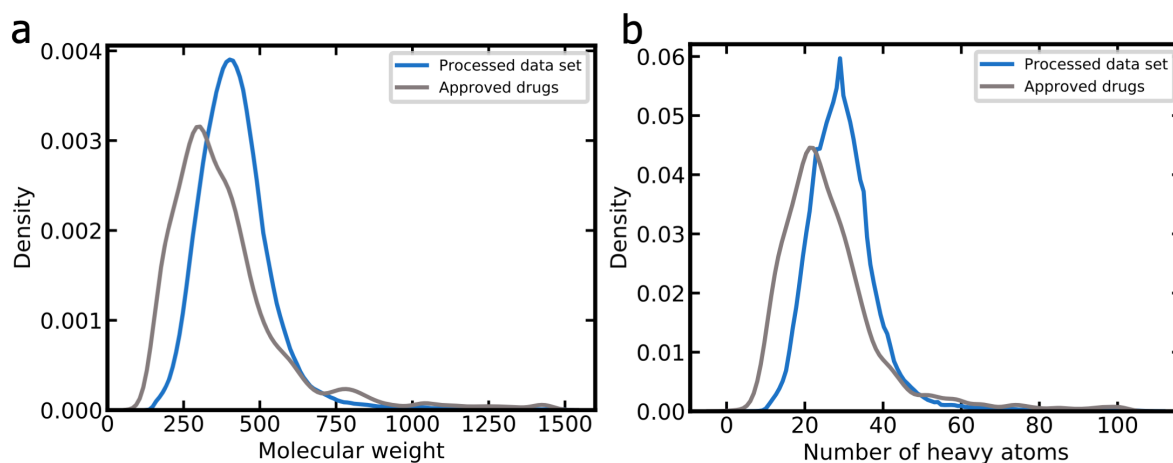


Figure S1. Density distribution of the (a) molecular weight and (b) the number of heavy atoms of all 481 194 compounds in the processed data set (blue; these are all valid compounds with at least one annotated bioactivity). The gray lines show the respective property distributions for the "Approved Drugs" subset of DrugBank¹ for reference.

REFERENCES

- (1) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.

Scientific Contributions

Publications

Publications of this dissertation

This overview summarizes the contributions of the author to the individual publications in scientific journals and a book chapter of this cumulative dissertation.

- D1 **Chen, Y.**; Kirchmair, J. Cheminformatics in Natural Product-Based Drug Discovery. *Mol. Inf.* **2020**, *39*, 2000171.

Y. Chen and J. Kirchmair conceptualized the work. Y. Chen analyzed the literature and wrote the largest part of the manuscript. J. Kirchmair supervised this work.

- D2 **Chen, Y.**; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* **2017**, *57* (9), 2099–2111.

Y. Chen, C. de Bruyn Kops and J. Kirchmair conceptualized the work. Y. Chen and C. de Bruyn Kops analyzed the literature and wrote the largest part of the manuscript. Y. Chen collected, curated and analyzed all the data. J. Kirchmair supervised this work.

- D3 **Chen, Y.**; de Bruyn Kops, C.; Kirchmair, J. Resources for Chemical, Biological, and Structural Data on Natural Products. In *Progress in the Chemistry of Organic Natural Products*; Kinghorn, A. D., Falk, H., Gibbons, S., Kobayashi, J., Asakawa, Y., Liu, J.-K., Eds.; Springer, 2019; Vol. 110, pp 37–71.

Y. Chen, C. de Bruyn Kops and J. Kirchmair conceptualized the work. Y. Chen analyzed the literature and collected, curated and analyzed all the data. Y. Chen wrote the largest part of the manuscript. J. Kirchmair supervised this work.

- D4 **Chen, Y.**; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58* (8), 1518–1532.

Y. Chen and J. Kirchmair conceptualized the work. Y. Chen analyzed the literature and collected, curated and analyzed all the data. This involved, among many other tasks, the validation and use of the tool “SugarBuster”, which was developed by M. Garcia de Lomana and N.-O. Friedrich, and validated by Y. Chen. Y. Chen wrote the largest part of the manuscript. J. Kirchmair supervised this work.

- D5 **Chen, Y.**; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **2019**, *9* (2), 43.

Y. Chen and J. Kirchmair conceptualized the work. Y. Chen collected, curated and prepared the data. Y. Chen developed the machine learning models and the web server, a task for which she received guidance from C. Stork. Y. Chen implemented the similarity maps, a task for which she received guidance from S. Hirte. Y. Chen wrote the largest part of the manuscript. J. Kirchmair supervised this work.

This publication has been selected by the editors as a *hot paper* (Editor's choice) of the journal *Biomolecules* in 2020.

- D6 **Chen, Y.**; Mathai, N.; Kirchmair, J. Scope of 3D Shape-Based Approaches in Predicting the Macromolecular Targets of Structurally Complex Small Molecules Including Natural Products and Macrocyclic Ligands. *J. Chem. Inf. Model.* **2020**, *60* (6), 2858-2875.

Y. Chen and J. Kirchmair conceptualized the work. Y. Chen collected, curated and analyzed all data. She also conducted all computational work. N. Mathai prepared and provided a high-quality data set from ChEMBL. Y. Chen wrote the largest part of the manuscript. J. Kirchmair supervised this work.

Additional publications

The following list itemizes the additional manuscripts authored by the PhD candidate that have been published in peer-reviewed journals to date.

- A1 Stork, C.; Embruch, G.; Šícho, M.; de Bruyn Kops, C.; **Chen, Y.**; Svozil, D.; Kirchmair, J. NERDD: A Web Portal Providing Access to in silico Tools for Drug Discovery. *Bioinformatics* **2020**, *36* (4), 1291-1292.
- A2 Stork, C.; **Chen, Y.**; Šícho, M.; Kirchmair, J. Hit Dexter 2.0: Machine-Learning Models for the Prediction of Frequent Hitters. *J. Chem. Inf. Model.* **2019**, *59* (3), 1030-1043.
- A3 Langeder, J.; Grienke, U.; **Chen, Y.**; Kirchmair, J.; Schmidtke, M.; Rollinger, J. M. Natural Products against Acute Respiratory Infections: Strategies and Lessons Learned. *J. Ethnopharmacol.* **2020**, *248*, 112298.
- A4 Mathai, N.; **Chen, Y.**; Kirchmair, J. Validation Strategies for Target Prediction Methods. *Briefings Bioinf.* **2019**, *21*(3), 791-802.
- A5 Xue, W.; Li, X.; Ma, G.; Zhang, H.; **Chen, Y.**; Kirchmair, J.; Xia, J.; Wu, S. N-Thiadiazole-4-Hydroxy-2-Quinolone-3-Carboxamides Bearing Heteroaromatic Rings as Novel Antibacterial Agents: Design, Synthesis, Biological Evaluation and Target Identification. *Eur. J. Med. Chem.* **2020**, *188*, 112022.

Oral Presentations

- O1 Oral presentation at 8th RDKit User Group Meeting. Applications of RDKit in Machine Learning. Sept 25-27, 2019, Hamburg, Germany
- O2 Oral presentation at 33rd Molecular Modelling Workshop Erlangen. NP-Scout: Machine Learning Models for the Identification and Visualization for the Natural Product-Likeness of Small Molecules. April 8-10, 2019, Erlangen, Germany
- O3 Invited talk at the Beijing University of Chemical Technology. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. Aug 31, 2017, Beijing, China

Poster Presentations

- P1 Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Identification of Natural Products and Natural Product-Like Compounds in Large Molecular Databases. 2nd RSC-BMCS/RSC-CICAG Artificial Intelligence in Chemistry organized by the Biological & Medicinal Chemistry Sector (BMCS) and Chemical Information & Computer Applications Sector (CICAG) of the Royal Society of Chemistry (RSC), Sept 2-3, 2019, Cambridge, UK
- P2 Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Readily Obtainable Natural Products Space. 22nd European Symposium on Quantitative Structure-Activity Relationships (EuroQSAR), Sept 16-20, 2018, Thessaloniki, Greece
- P3 Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. 11th International Conference on Chemical Structures (ICCS), May 27-31, 2018, Noordwijkerhout, The Netherlands
- P4 Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Comparative Analysis of the Chemical Space of Known and Purchasable Natural Products. 32nd Molecular Modelling Workshop (MMWS), Mar 12-14, 2018, Erlangen, Germany
- P5 Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Analysis of Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. EUROPIN Summer School on Drug Design, Sept 17-22, 2017, Vienna, Austria

- P6 Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Analysis of Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. 2017 Chinese Academic Conference on Medicinal Chemistry (CMCS) and 2017 Chinese Pharmaceutical Association-the European Federation for Medicinal Chemistry International Symposium on Medicinal Chemistry (CPA-EFMC ISMC), Aug 27-30, 2017, Beijing, China

Awards

Poster Prize at 2nd RSC-BMCS/RSC-CICAG Artificial Intelligence in Chemistry organized by the Biological & Medicinal Chemistry Sector (BMCS) and Chemical Information & Computer Applications Sector (CICAG) of the Royal Society of Chemistry (RSC), Sept 2-3, 2019, Cambridge, UK