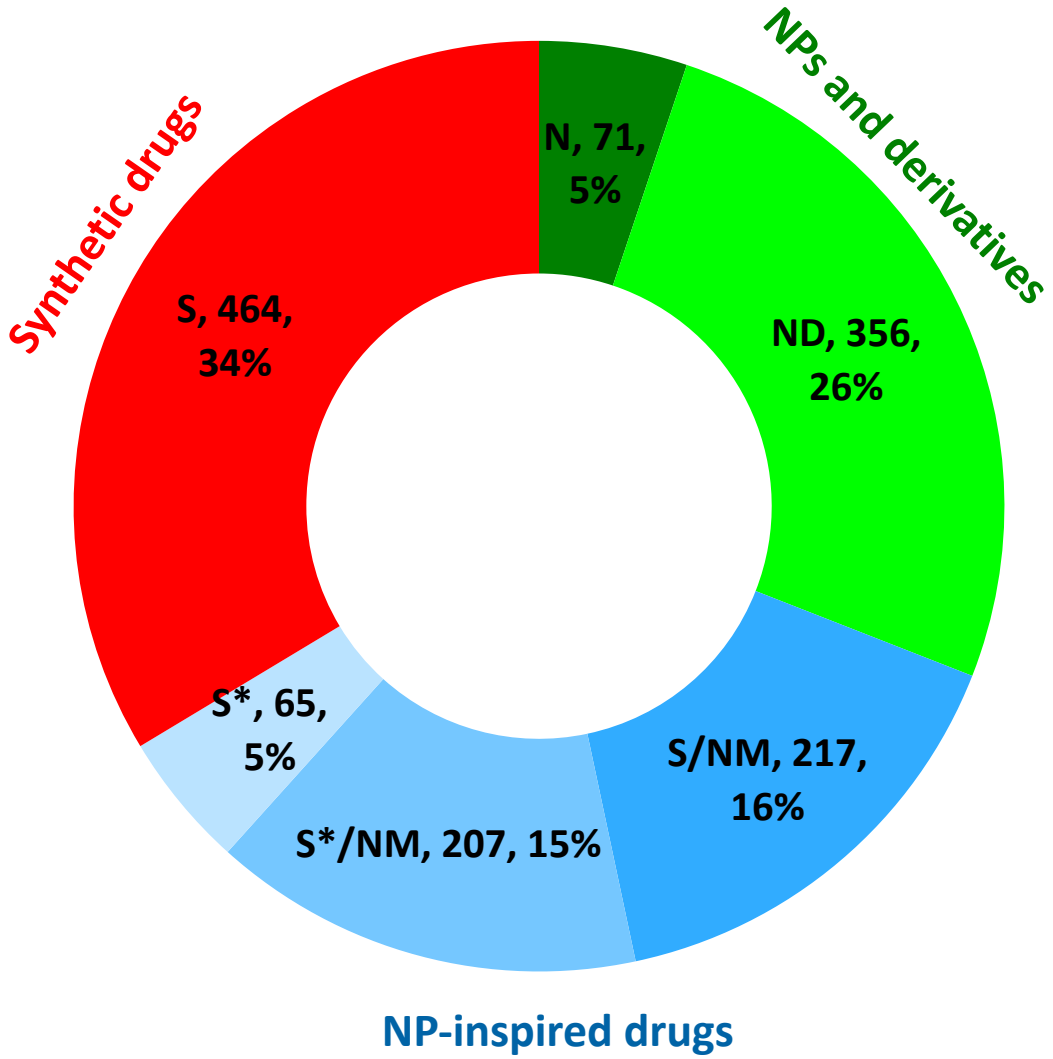
universität wien

Chemoinformatics in Natural Product-Based Drug Discovery

Johannes Kirchmair

Slides available
upon request and from:

Sources for small-molecule drugs (1981-2019)



More than half of all modern small-molecule drugs are linked to natural products.

However, a closer look reveals that ***many of these compounds either are of high molecular weight (typically peptides) or members of few prominent scaffolds (e.g. steroids or macrolides such as rapamycin analogs)***

N	natural product
ND	natural product derivative (usually semi-synthetic)
S	synthetic drug
S*	synthetic drug with natural product pharmacophore
/NM	mimic of natural product

Newman D and Cragg G, J Nat Prod 2020 , 83, 770−803.

Bioactivity of natural products

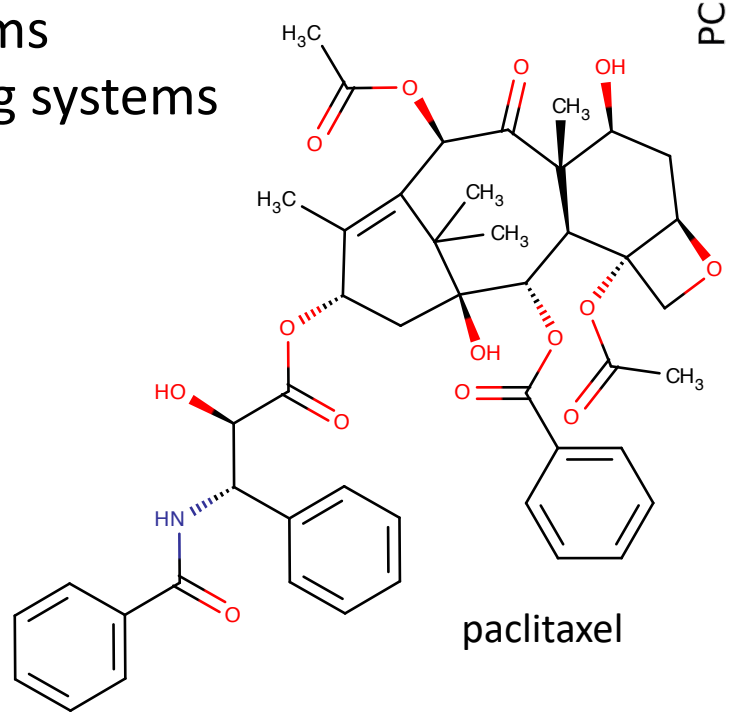
- **Plants can't run away!** They need to protect themselves from fungal infections, caterpillars, sheep, cattle, etc.
  → They require substances that trigger a biological response in the target organism: **bitter, hot or toxic**

- Natural products have been undergoing **optimization throughout the course of evolution**
  → **Wide range of biological activities in different organisms**
  → "**Privileged scaffolds**"
  → **Higher hit rates** in biological assays than traditional synthetic libraries

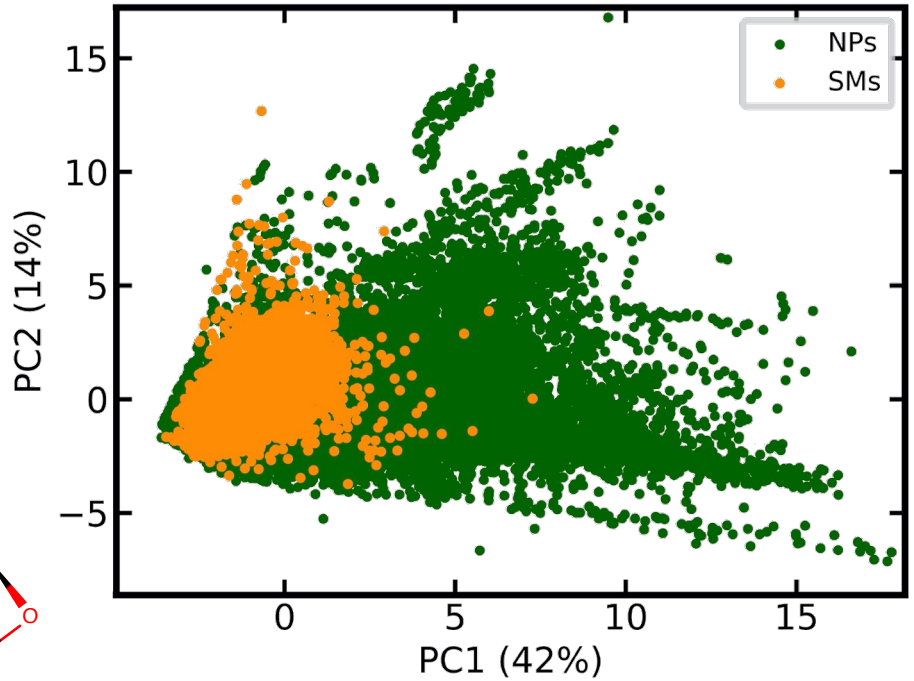
- Protein structures from different organisms are often similar
  → what works in plants may also trigger biological effects in, e.g., humans
  - On the contrary: huge differences in ADME between species

Hallmarks of natural products

- Enormous physicochemical and structural diversity
- In part high molecular complexity:
  - Higher molecular weight
  - High 3D shape complexity; fewer planar fragments
  - High number of sp^3-hybridized carbon atoms, including also bridgehead atoms
  - More complex and diverse ring systems
  - Sugar moieties



paclitaxel



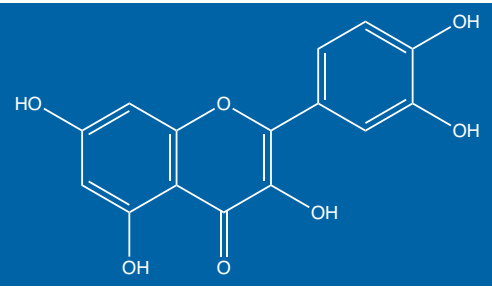
 Chen Y. et al., Biomolecules 2019, 9, 43.

Challenges in natural products drug discovery

- **Limited availability/costs** of materials for testing
- Difficulties in harvesting, transport (also w.r.t. import and export of natural products), isolation, testing and synthesis
- Problems related to decomposition, aggregation, precipitation and reactivity:
  high false-positive hit rates in biological assays
- Computational methods:
  - In part high **molecular complexity and flexibility**
  - 3D in silico approaches depend on molecular structures with correct **stereochemistry**
  - Many computational methods have been **designed for synthetic drug-like molecules** rather than natural products. Modifications for use with natural products may be required

universität wien

Is quercetin a “super drug”?



As of today, the PubChem Bioassay database lists “conclusive” testing results for quercetin for 427 proteins, with quercetin reported as active on 268 of these proteins

(and most of these proteins being of pharmaceutical relevance)

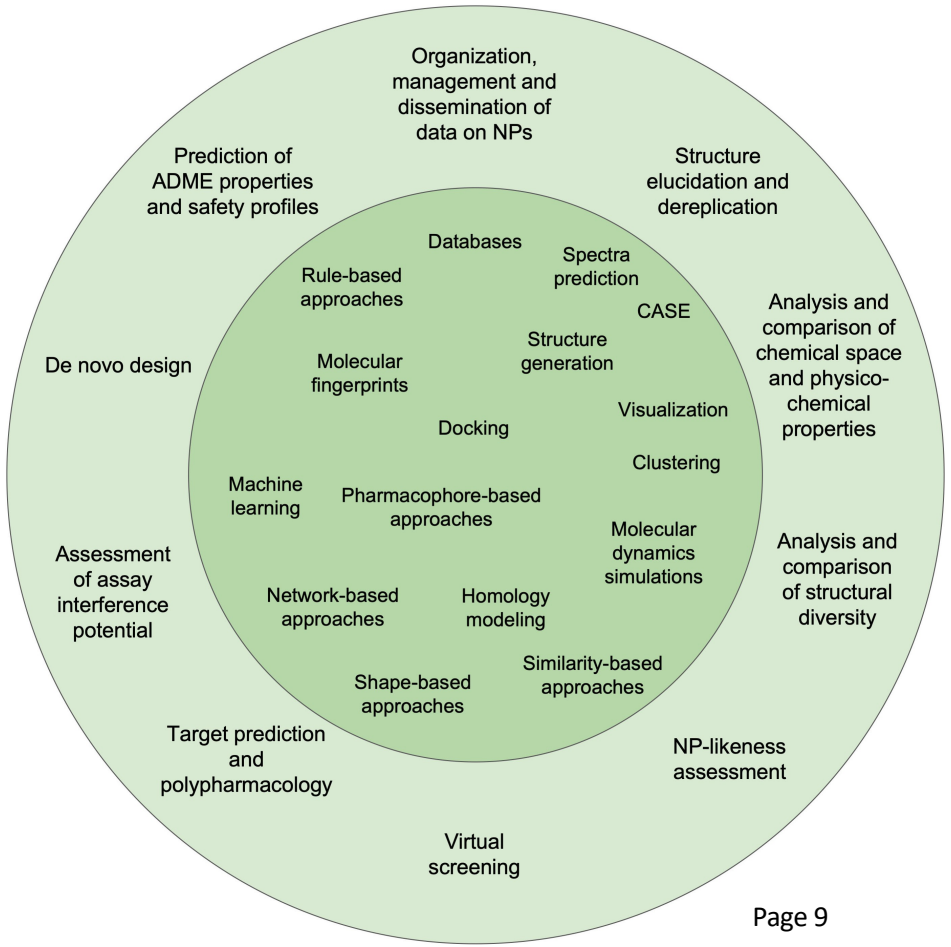
Cheminformatics in natural product-based drug discovery

- **Retrieval, organization, curation and management** of chemical information
- **Visualization and analysis**
- **Natural product dereplication**
- **Molecular property prediction**
- **Virtual screening**: identification of bioactive natural products and promising **species**
- Prediction of the **bioactivity spectra** and, hence, the biomacromolecular targets of natural products
- Prediction of **ADME properties and toxicity of natural products**
- **Design** of nature-inspired compounds
- **NP-likeness** assessment
- Identification **bioactive natural products** (virtual screening)

Review www.molinf.com molecular informatics
DOI: 10.1002/minf.202000171

Cheminformatics in Natural Product-based Drug Discovery

Ya Chen[a] and Johannes Kirchmair*[a, b]



Saldívar-González F I et al., Chem Sci 2022, 13, 1526-1546.
Schneider P et al., Chimia 2022, 76, 396.
Chen Y and Kirchmair J, Mol Inf 2020, 39, 2000171.

Identification bioactive natural products (virtual screening)

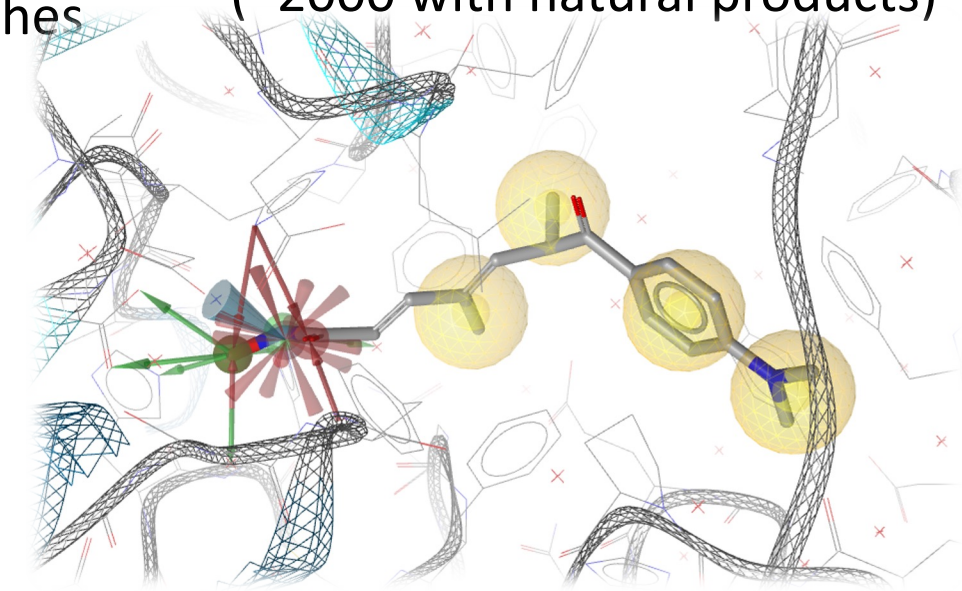
Virtual screening is established as an important tool in computer-guided natural products research

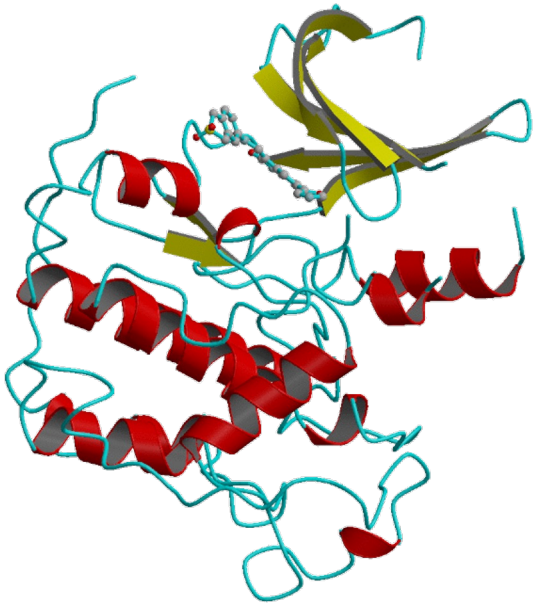
Established virtual screening methods

- 2D and 3D similarity-based approaches
- Reduced graph representations
- Pharmacophore models
- Molecular shape-based approaches
- Ligand docking methods
- Machine learning models

Challenges

- Structural complexity (conformational flexibility; stereochemistry)
- Limited quality, coverage and size of bioactivity datasets
- Limited availability of protein structural data (~2000 with natural products)





Encyclopedic and general natural products databases

Data source name	Scope	Number of compounds	Biological data	Free use	Bulk data access	Chemistry-aware web interface
COCONUT	All forms of life	>400k	none	yes	yes	yes
SuperNatural II	All forms of life	>325k	bioactivity and toxicity data	yes	no	yes
Reaxys	All forms of life	>260k	bioactivity data	no	yes	yes
Dictionary of Natural Products (DNP)	All forms of life	>230k	bioactivity data	no	yes	yes
UNPD	All forms of life	>229k	none	yes	no longer avail.	no
AntiBase	Microorganisms and higher fungi	> 43k	bioactivity data (focus on antimicrobial activity)	no	no	yes
CMAUP	Plants	>47k	bioactivity data	yes	yes	yes
NPASS	All forms of life	~35k	bioactivity data	yes	no	yes
The Natural Products Atlas	Bacteria and fungi	>32k	none	yes	yes	yes
Pye et al. data set	NPs from microorganisms and marine life published between 2012 and 2015	> 6k	none	yes	yes	no
Natural products included in the PubChem Substance Database	All forms of life	>3.5k	bioactivity data	yes	yes	yes
UEFS Natural Products	None specified	~500	none	via ZINC	via ZINC	no

Chen Y. et al., J Chem Inf Model 2017, 57, 2099−2111.
Chen Y. et al., In *Progress in the Chemistry of Organic Natural Products 110* (2019): Cheminformatics in Natural Product Research. ISBN 978-3-030-14632-0.

NP databases focused on traditional medicines

Data source name	Scope	Number of compounds	Biological data	Free use	Bulk data access	Chemistry-aware web interface
TCM database@Taiwan	Chinese medicinal herbs	>60k	bioactivity data	yes	yes	yes
TCMID 2.0	Chinese medicinal herbs	>43k	bioactivity data	yes	yes	no
YaTCM	Chinese medicinal herbs	>47k	bioactivity data	yes	no	yes
Chem-TCM	Chinese medicinal herbs	>12k	bioactivity data	no	yes	no
HIM	Chinese medicinal herbs	~1300	ADME and toxicity data	yes	via ZINC	via ZINC
HIT	Chinese medicinal herbs	~530	bioactivity data	yes	via ZINC	via ZINC
IMPPAT	Indian medicinal herbs	>9500	bioactivity data	yes	no	yes

Chen Y. et al., J Chem Inf Model 2017, 57, 2099−2111.
Chen Y. et al., In *Progress in the Chemistry of Organic Natural Products 110* (2019): Cheminformatics in Natural Product Research. ISBN 978-3-030-14632-0.

NP Databases focused on a specific habitat or geographical region

Data source name	Scope	Number of compounds	Biological data	Free use	Bulk data access	Chemistry-aware web interface
DMNP	Marine life	>55k (including NP derivatives)	bioactivity data	no	no	yes
MarinLit	Marine life	>33k	bioactivity data	no	no	yes
TIPdb	Taiwanese herbs	~9000	bioactivity data (focus on anticancer, antiplatelet and antituberculosis activity)	yes	yes	no
NANPDB	All forms of life indigenous to North Africa	>6800	bioactivity data	yes	yes	yes
AfroDb	African medicinal plants	~1000	bioactivity data	yes	yes	no
SANCDB	South African plants and marine life	>700	none	yes	yes	yes
AfroCancer	African medicinal plants with confirmed anticancer, cytotoxic or antiproliferative activity	~400	bioactivity data (focus on anticancer activity)	yes	yes	no
AfroMalariaDB	African plant NPs with confirmed antimalarial or antiplasmodial activity	>250	bioactivity data (focus on antimalarial activity)	yes	yes	no
$NuBBE_{DB}$	NPs from Brazilian plants, fungi, insects, marine organisms, and bacteria	>2200	bioactivity data (focus on antimicrobial activity)	yes	yes	yes
BIOFACQUIM	NPs from plants, fungi and propolis isolated and characterized in Mexico	>400	bioactivity data	yes	yes	no

Chen Y. et al., J Chem Inf Model 2017, 57, 2099−2111.
Chen Y. et al., In *Progress in the Chemistry of Organic Natural Products 110* (2019): Cheminformatics in Natural Product Research. ISBN 978-3-030-14632-0.

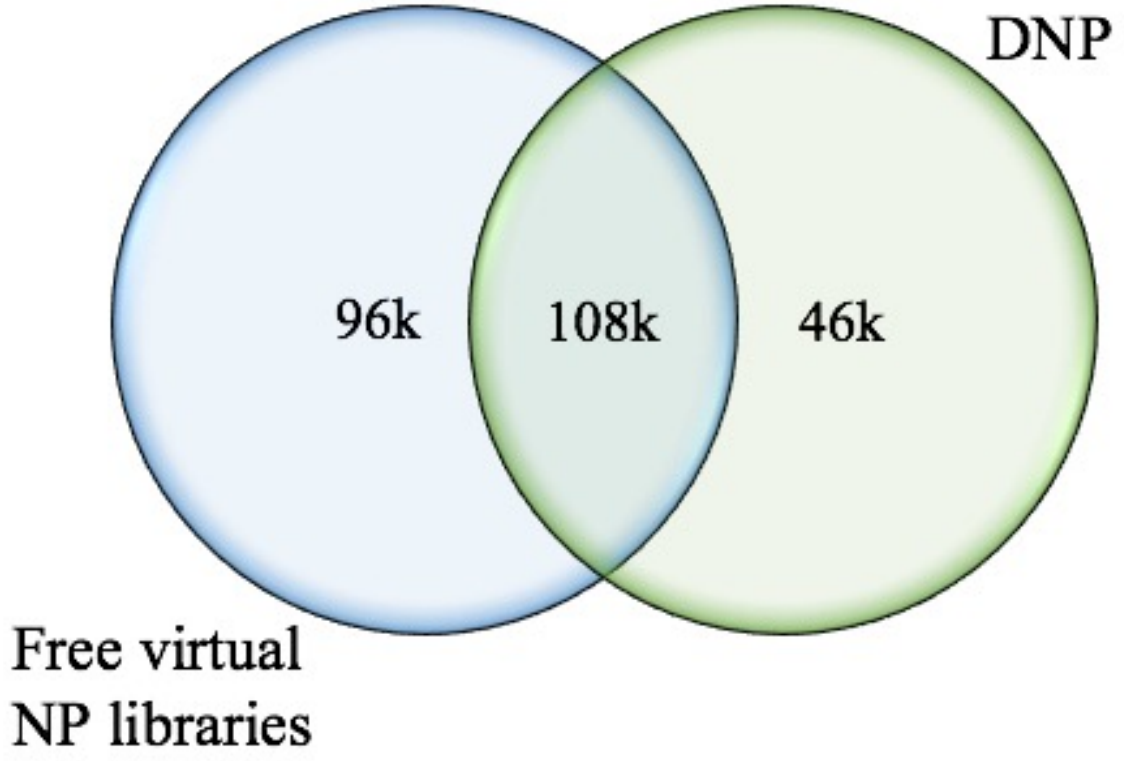
Further NP databases

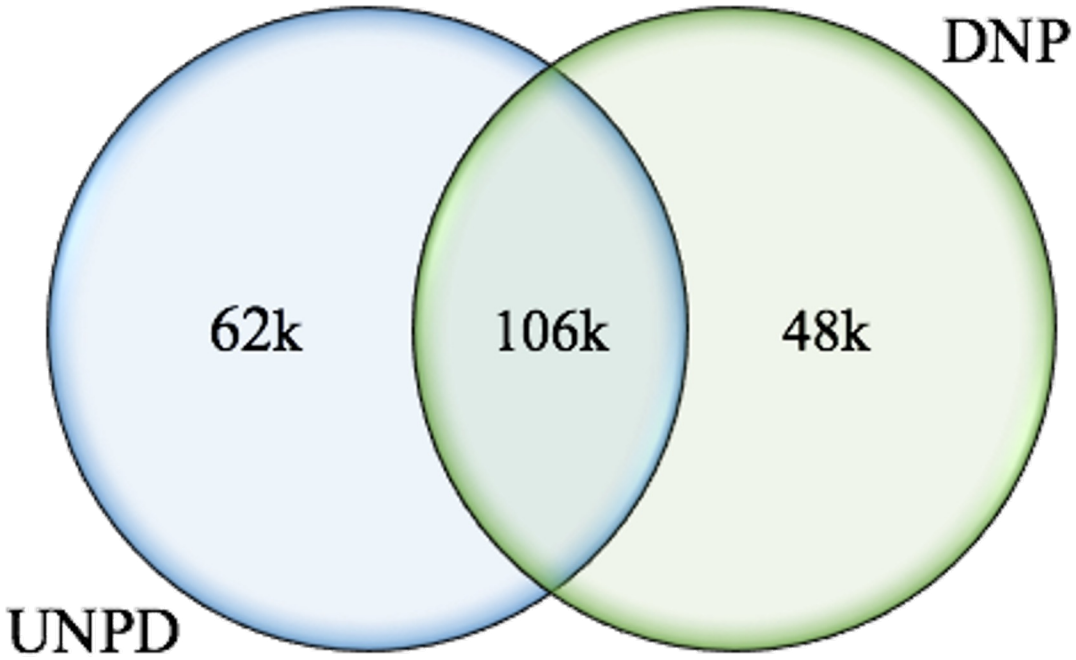
Data source name	Scope	Number of compounds	Biological data	Free use	Bulk data access	Chemistry-aware web interface
Databases focused on specific organisms						
PAMDB	*Pseudomonas aeruginosa*	>4300	bioactivity data	yes	yes	yes
StreptomeDB 2.0	Streptomycetes	~4000	bioactivity data	yes	yes	yes
Databases focused on specific biological activities						
NPCARE	NPs with measured anticancer activity, sourced from plants, marine species and microorganisms	>6500 >1500 in bulk download	bioactivity data (focus on anticancer activity)	yes	yes	no
NPACT	NPs with measured anticancer activity, sourced from plants	>1500	bioactivity data (focus on anticancer activity)	yes	via ZINC	yes
InflamNat	NPs with measured antiinflammatory activity, sourced primarily from terrestrial plants	>650	bioactivity data (focus on antiinflammatory)	yes	yes	no
Databases focused on specific natural product classes						
Carotenoids Database	Carotenoids extracted from almost 700 source organisms	>1100	bioactivity data	yes	no	yes

Chen Y. et al., J Chem Inf Model 2017, 57, 2099−2111.
Chen Y. et al., In *Progress in the Chemistry of Organic Natural Products 110* (2019): Cheminformatics in Natural Product Research. ISBN 978-3-030-14632-0.

Free vs. commercial virtual natural product libraries

Structures of more than 250k NPs have been deposited to date



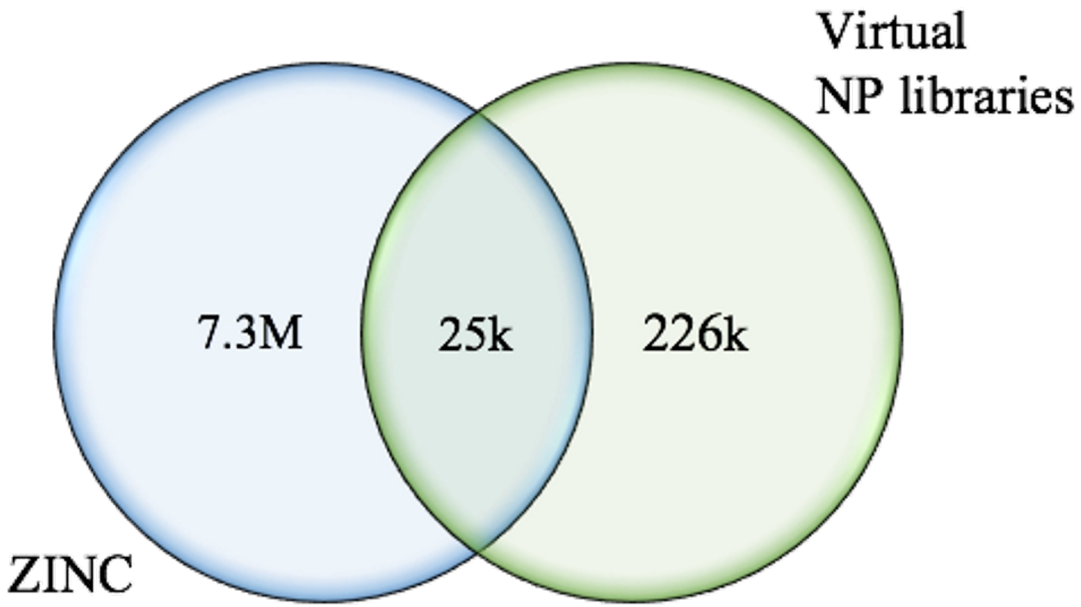


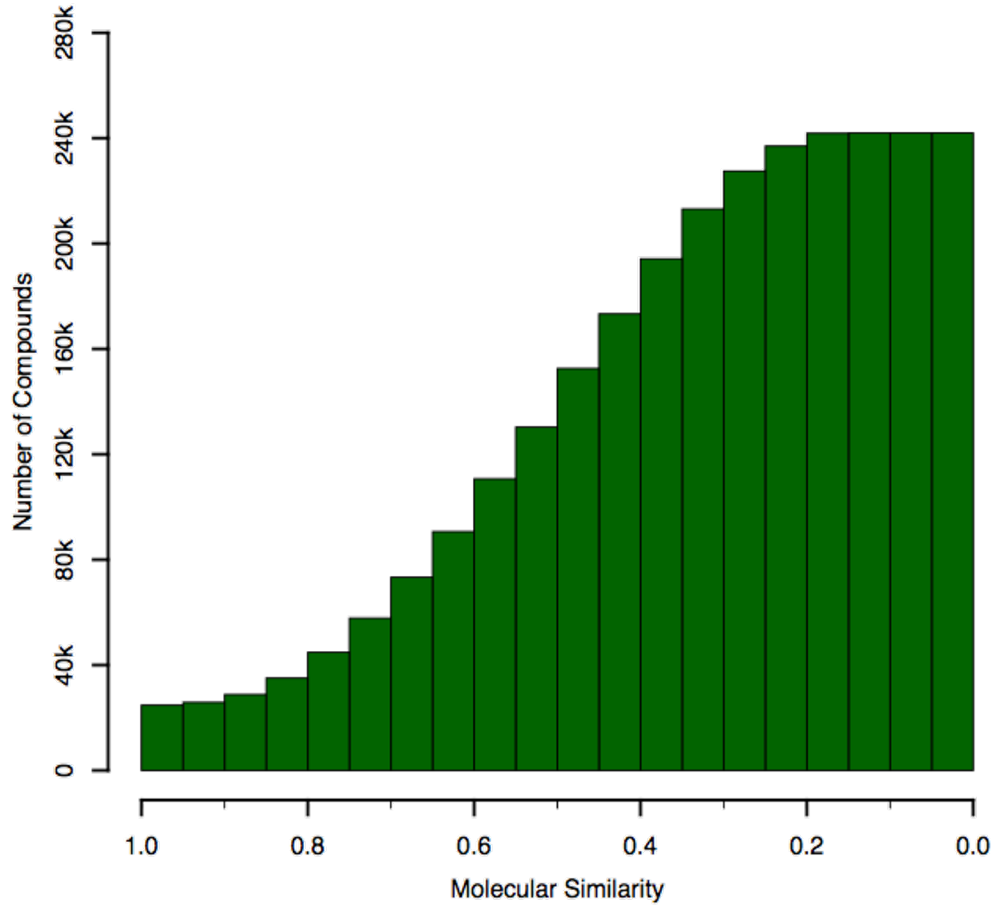
- 70% of the compounds in the DNP can be found in at least one free library
- 53% of all compounds contained in free libraries are also covered by the DNP

- UNPD is one of the largest free sources of chemical data on NPs
- Roughly two-thirds of DNP and UNPD data overlap

 Chen Y. et al., J Chem Inf Model 2017, 57, 2099−2111.

Readily obtainable NPs and derivatives





- Significant effort involved in sourcing natural products
- Only approx. 25k (10%) of all known NPs are readily obtainable
- At a Morgan2 fingerprint-based Tanimoto coefficient of 0.7, about 25% of all known NPs are covered by obtainable NPs and their analogs

 Chen Y. et al., J Chem Inf Model 2017, 57, 2099−2111.

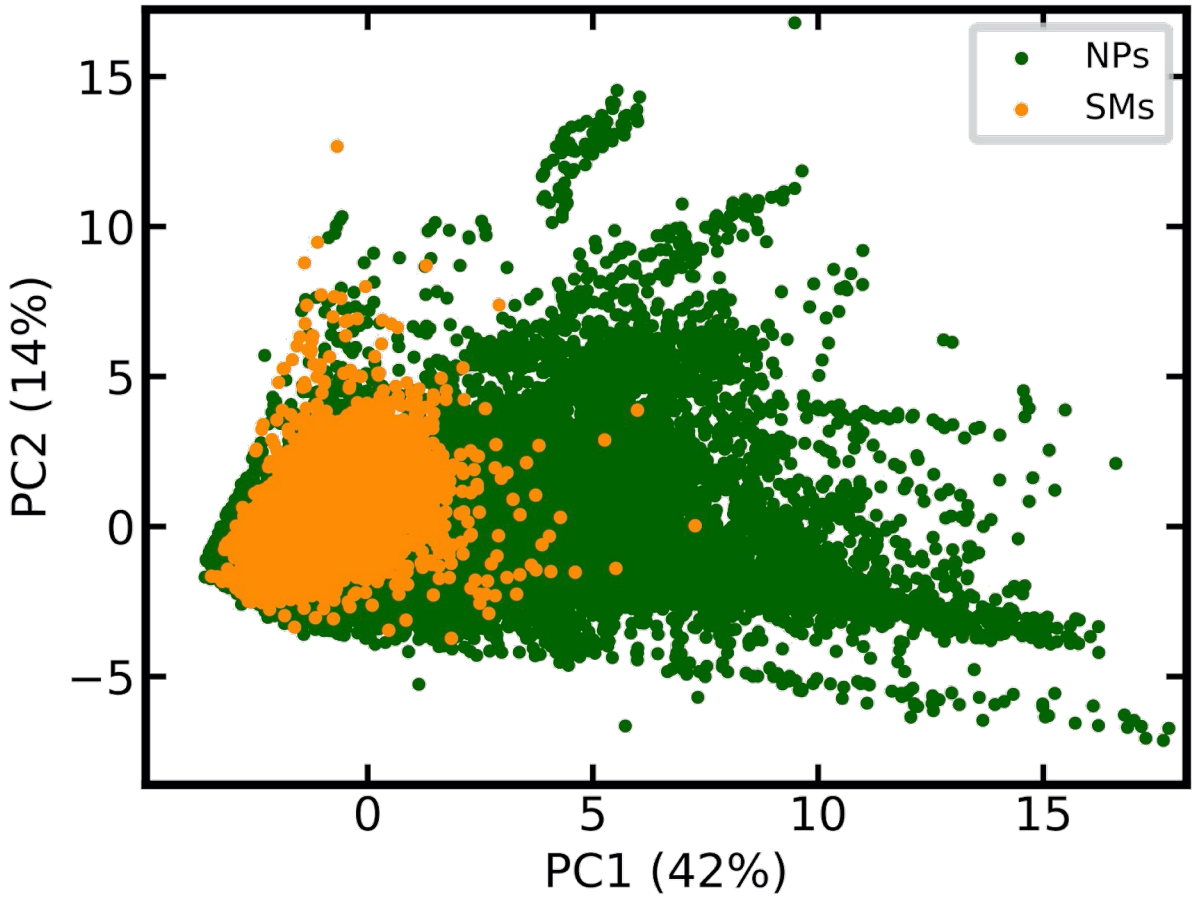
Natural products readily purchasable from suppliers

No. 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 Chemhere, Oxchem Corporation, BePharm, Zelinsky Institute, Combi-Blocks, Debye Scientific, Matrix Scientific, WuXi AppTec, Ark Pharm, Bide Pharmatech, BioSynth, InterBioScreen, Labseeker, StruChem, Alfa-Aesar
2000–3000	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, 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
100–1000	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

 Chen Y. et al., J Chem Inf Model 2017, 57, 2099−2111.

NP-Scout: development of a method for the assessment of natural product-likeness

- Use cases:
  - Profiling of databases (% NPs; NP-likeness)
  - Identification of genuine NPs in commercial compound libraries, which often contain also synthetic molecules
  - Prioritization of compounds for experimental testing
  - Library design
- Data set:
  - 201 761 unique NPs (multiple free sources)
  - 201 761 unique synthetic compounds (ZINC)



 Chen Y. et al., Biomolecules 2019, 9, 43.

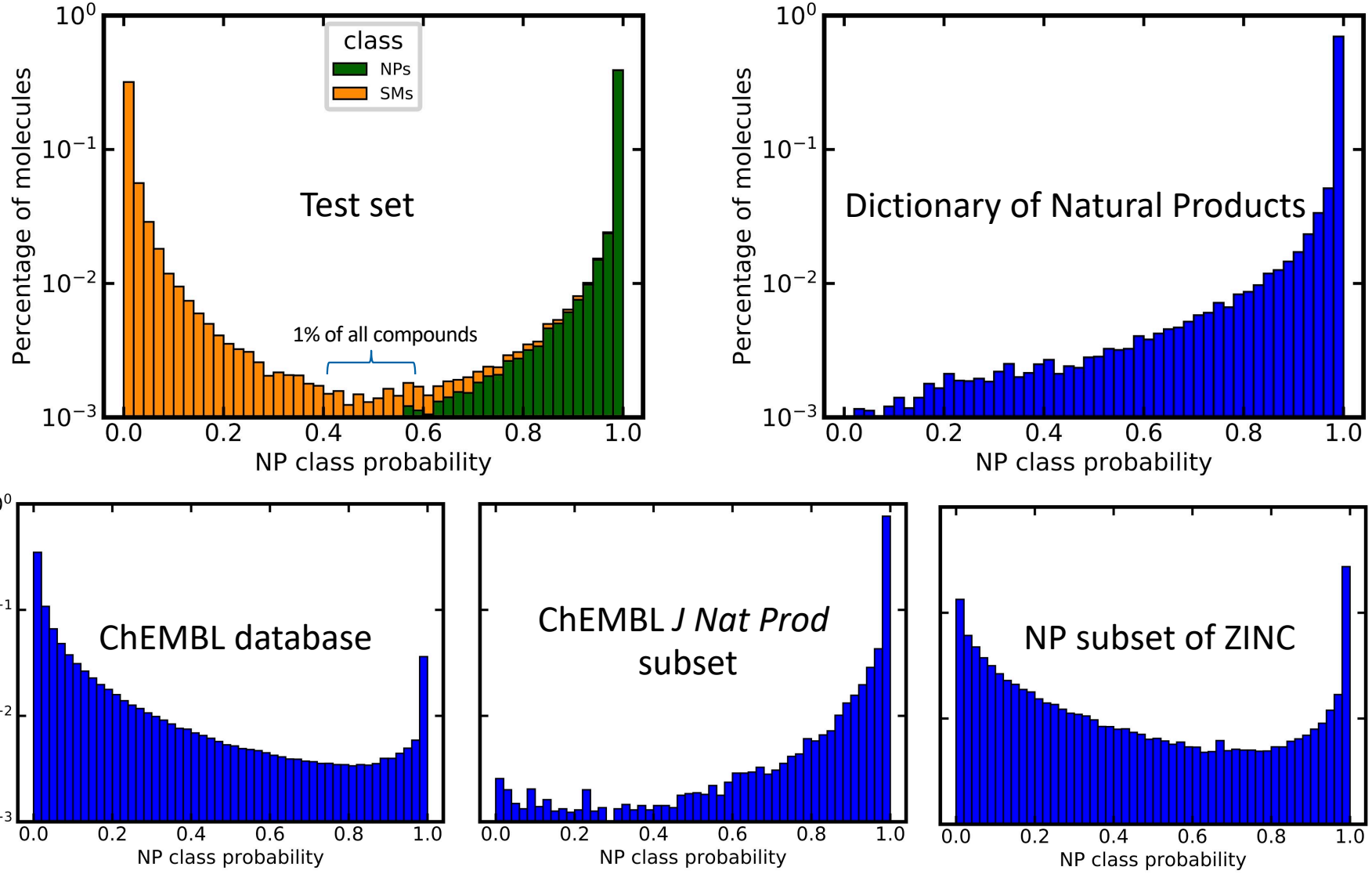
NP-Scout: Modelling approach

- Machine learning approach
  - Random forest
- Descriptors:
  - MACCS keys
  - Morgan2 fingerprints
  - (MOE 2D descriptors)

Test	Metric	MOE 2D descriptors	Morgan2 fingerprints (1024 bits)	MACCS keys	NP-Likeness calculator
10-fold cross-validation	**AUC**	0.997	0.997	0.997	/
	MCC	0.953	0.959	0.959	/
Holdout data	**AUC**	0.997	0.997	0.997	0.997
	MCC	0.954	0.960	0.960	0.959

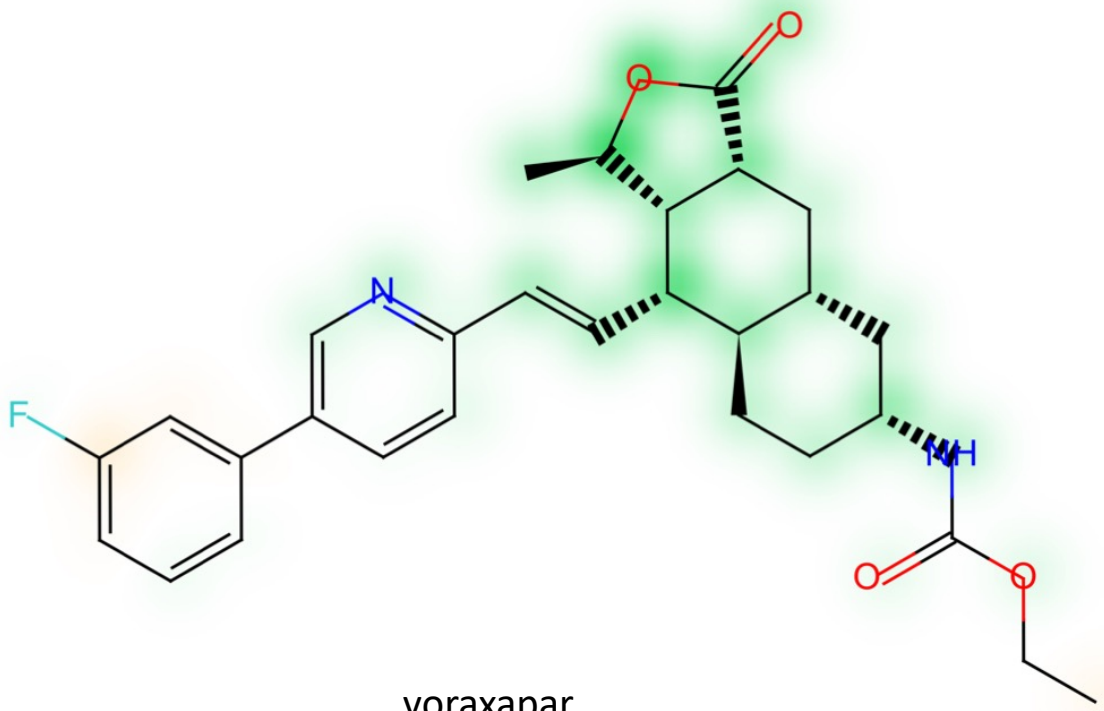
Chen Y. et al., Biomolecules 2019, 9, 43.
Ertl et al., J Chem Inf Model 2008, 48, 68-74.

NP-Scout: Model validation

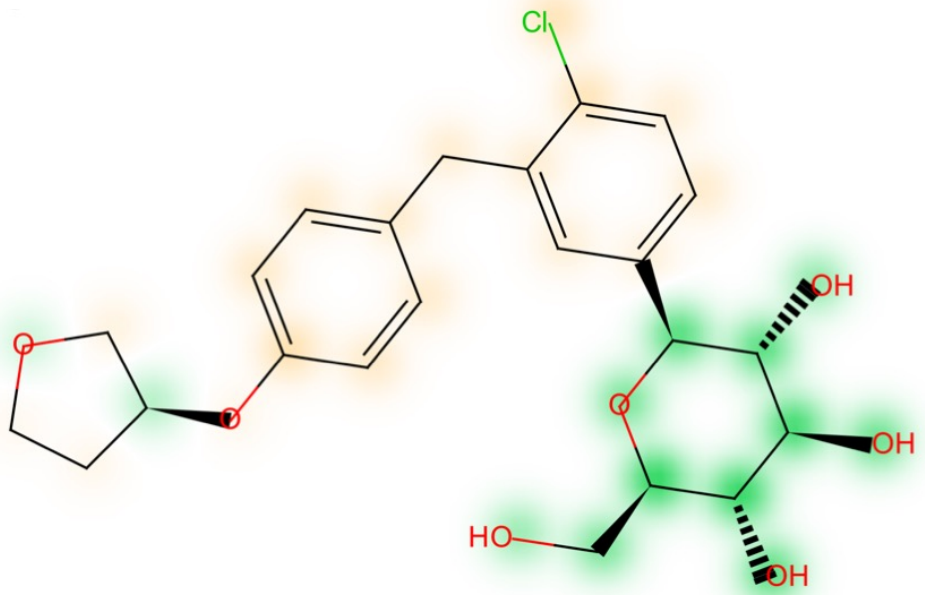


 Chen Y. et al., Biomolecules 2019, 9, 43.

NP-Scout: Similarity maps



voraxapar



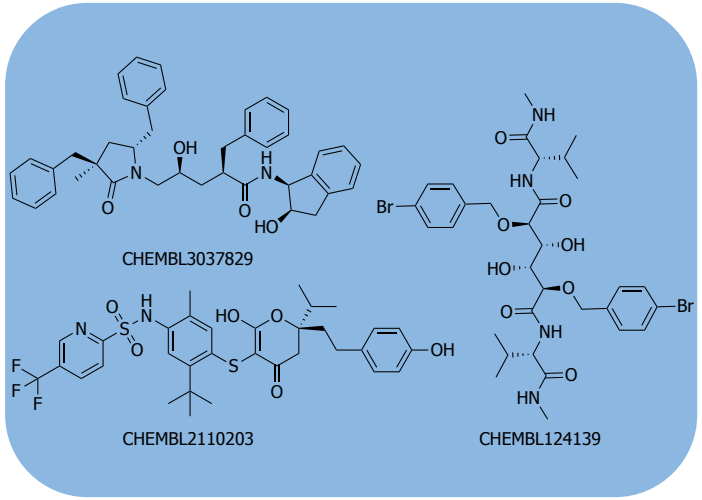
empagliflozin

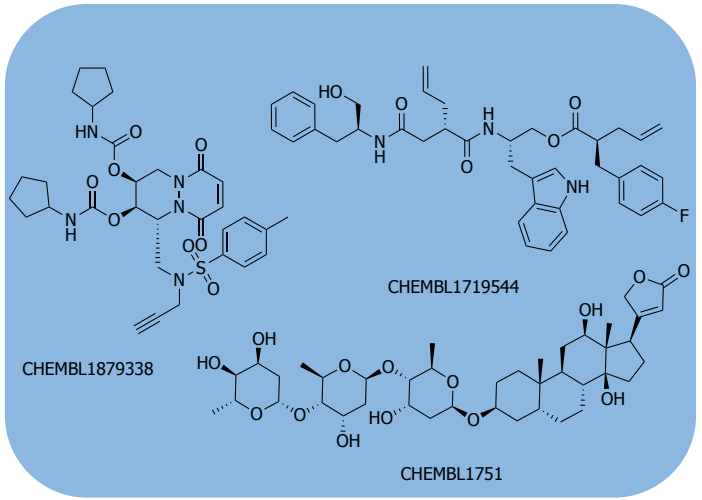
Target prediction methods based on 3D molecular shape similarity: How far can we get?

HIV-1 protease

Paired box protein Pax-8

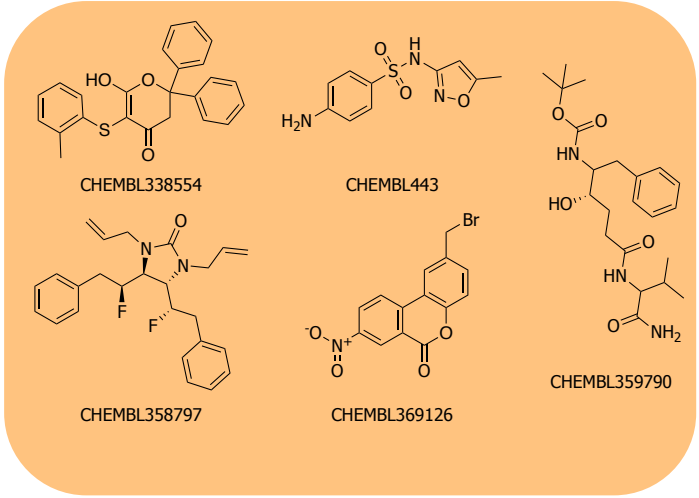
Complex small-molecules

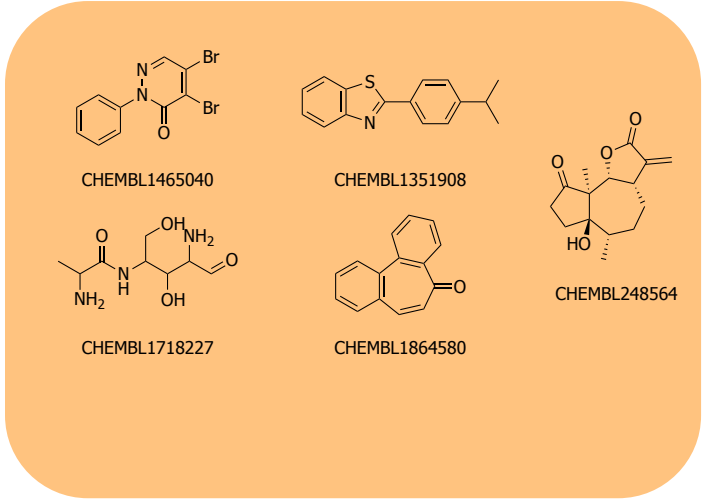




Any compounds consisting of at least 45 heavy atoms
(631 Da on average)
or macrocyclic with at least 30 heavy atoms
(772 Da on average)

Non-complex small-molecules

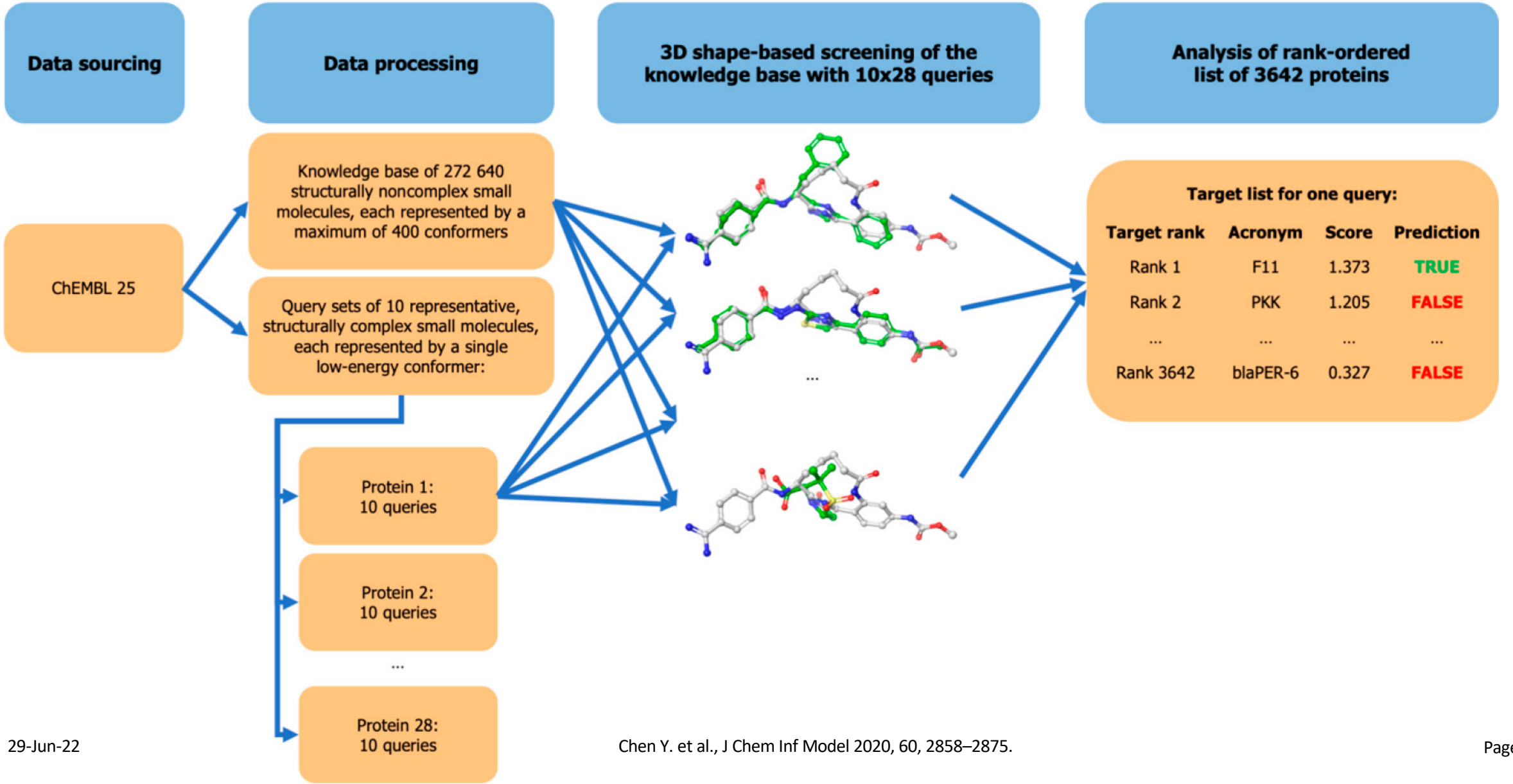




Any compounds consisting of 15–30 heavy atoms
(222 to 424 Da on average)

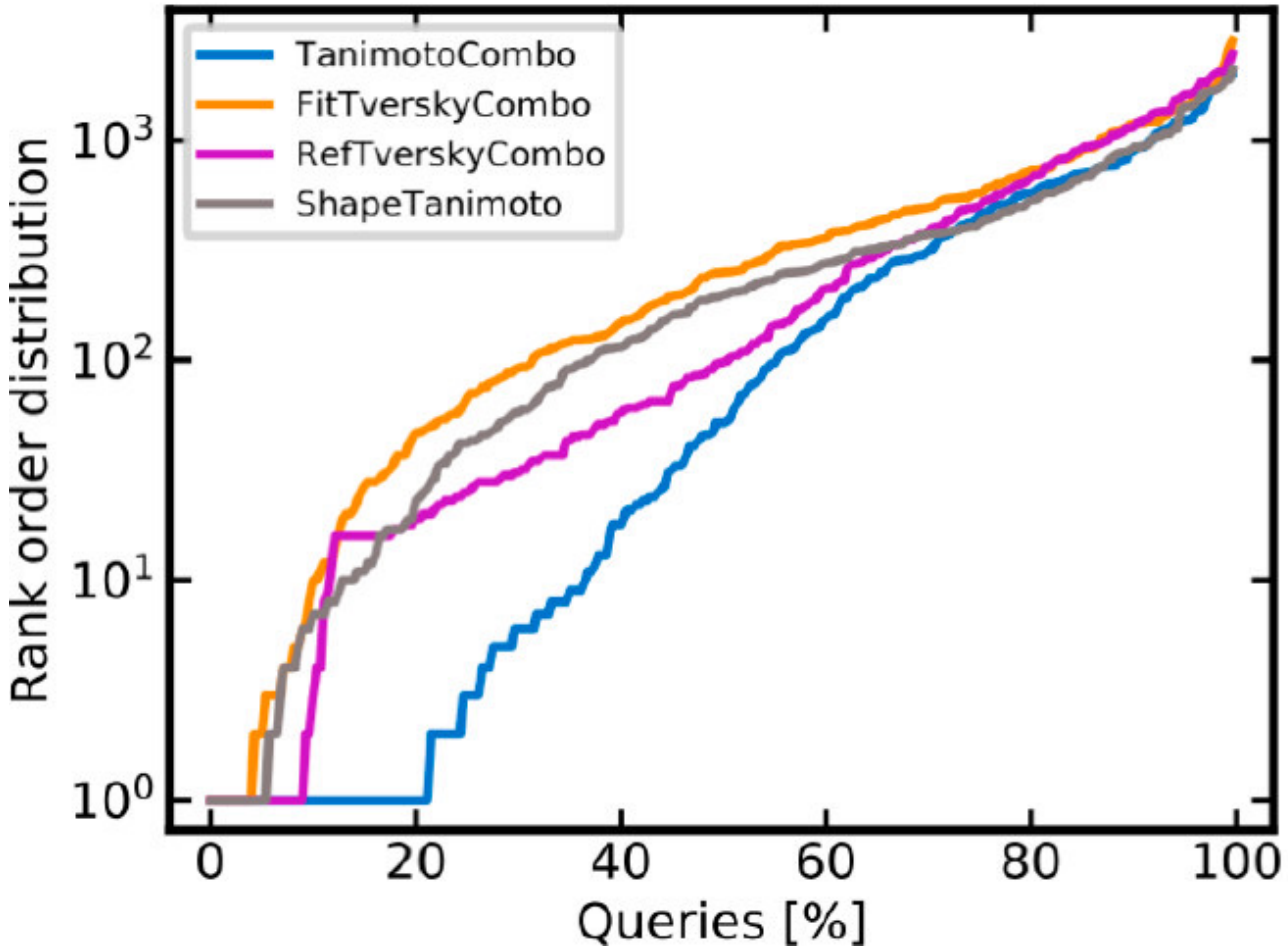
Chen Y. et al., J Chem Inf Model 2020, 60, 2858–2875.

Target prediction methods based on 3D molecular shape similarity: How far can we get?



Chen Y. et al., J Chem Inf Model 2020, 60, 2858–2875.

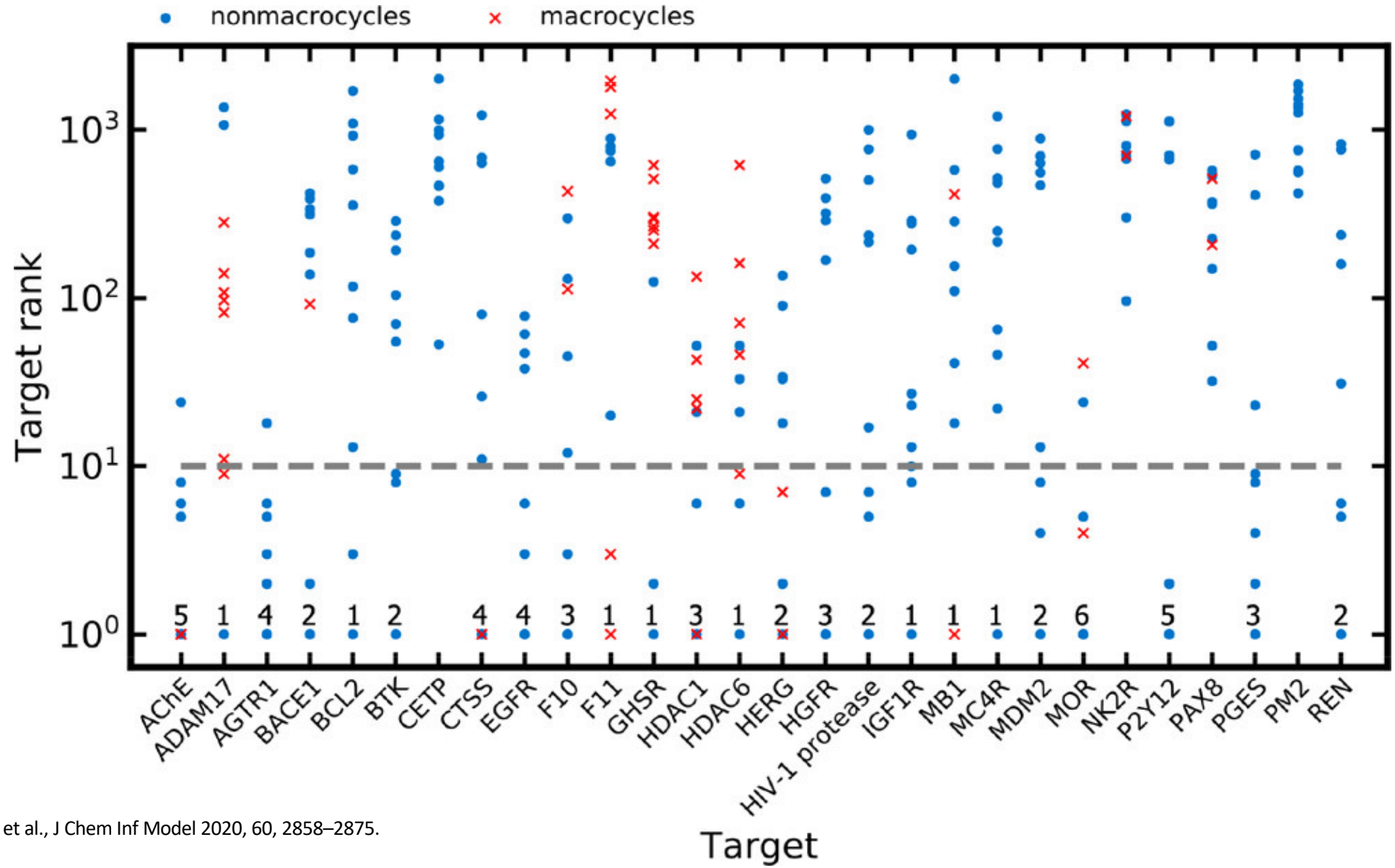
Ranking performance



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 all queries.

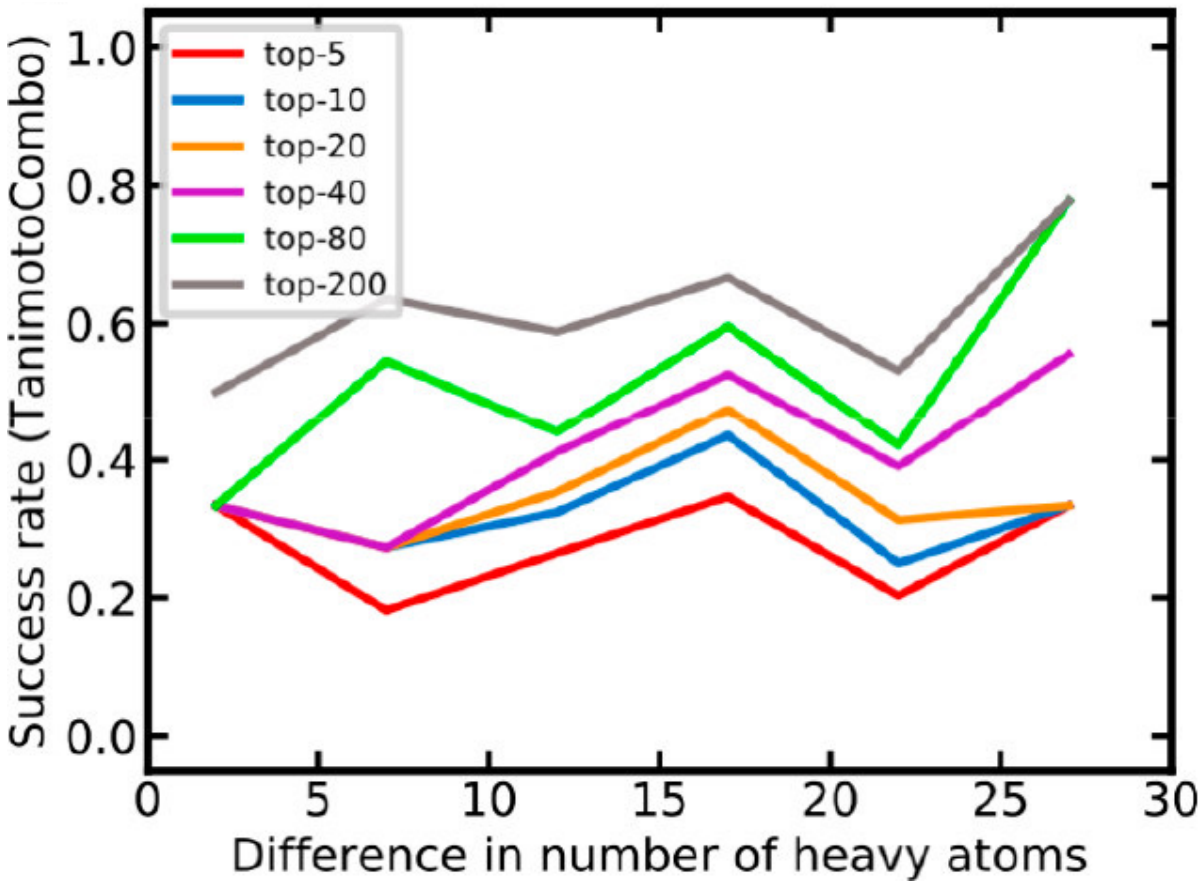
 Chen Y. et al., J Chem Inf Model 2020, 60, 2858–2875.

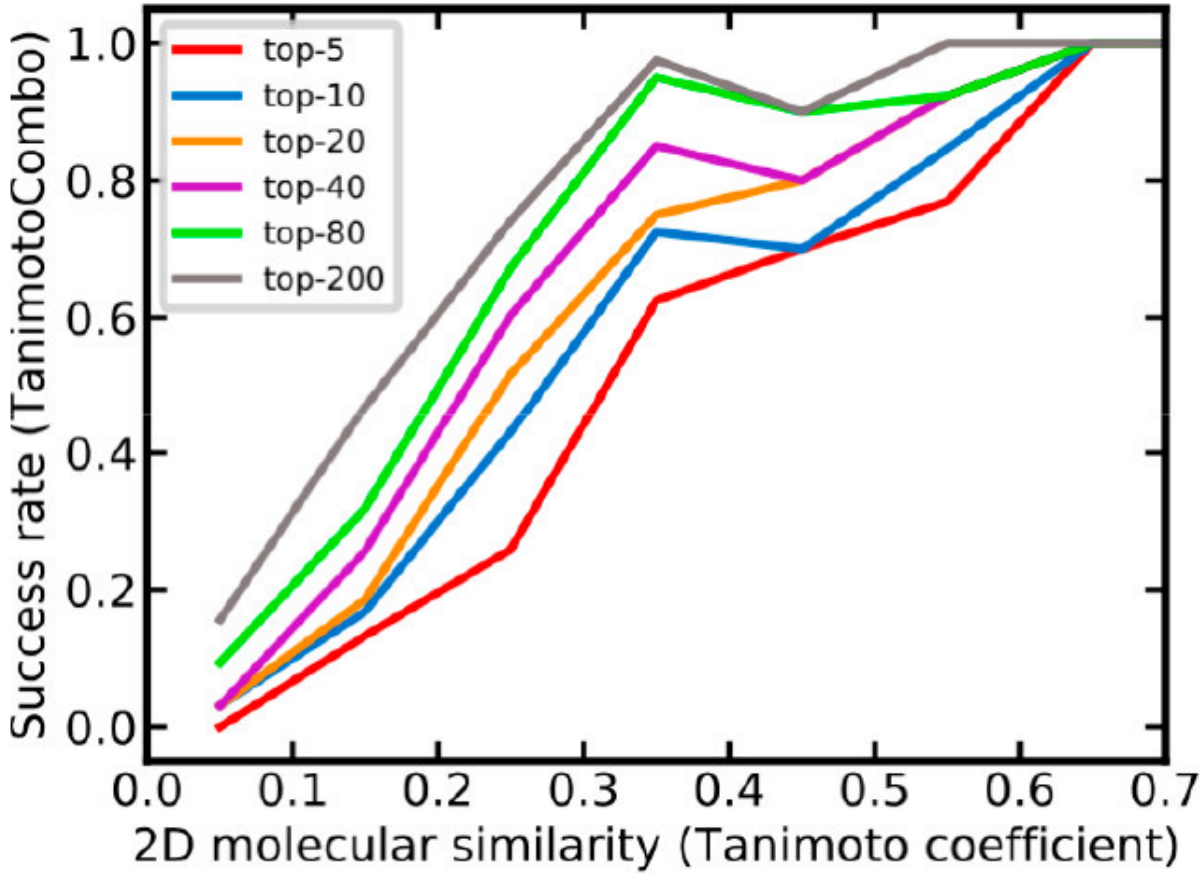
Ranks assigned with the TanimotoCombo score to the target of interest for the 280 complex small-molecule queries



 Chen Y. et al., J Chem Inf Model 2020, 60, 2858–2875.

Performance as a function of molecular similarity



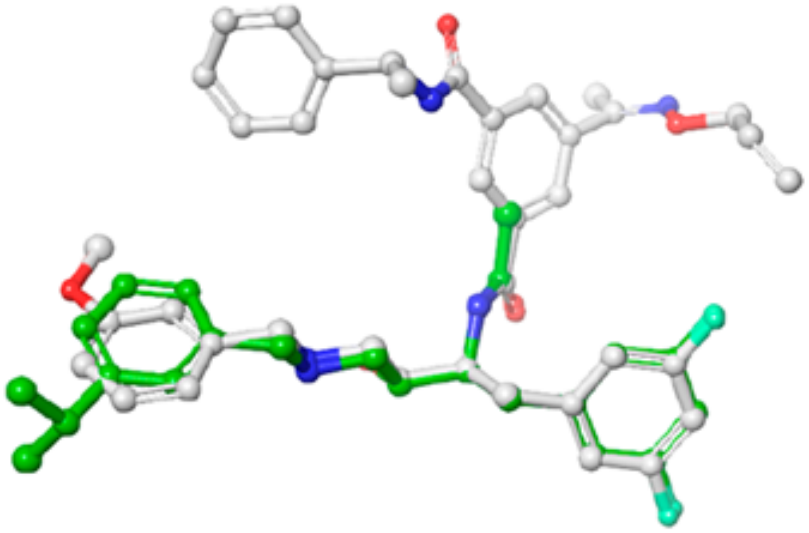


Success rate = fraction of complex small molecule queries ranked among the top-*k* positions

 Chen Y. et al., J Chem Inf Model 2020, 60, 2858–2875.

Lessons learned on in silico target prediction

- Method ranked the targets of roughly **one-third of the 280 complex small-molecules among the top-5 positions of a list of 3642 proteins**
- Methods based on 3D molecular shape comparison hold promise to **identify distant similarity**
- **ROCS produces "tidy" overlays** that help researchers to judge the plausibility of individual predictions → even bad matches can be valuable information (indication of novelty, etc.)
- **Final prediction relies on a single data point** for which validity can be checked individually →
  working with noisy data less problematic
- Even these methods are **challenged by macrocyclic compounds and natural products** (the available data are clearly limited)
- These and other target prediction methods should always be worth a try!



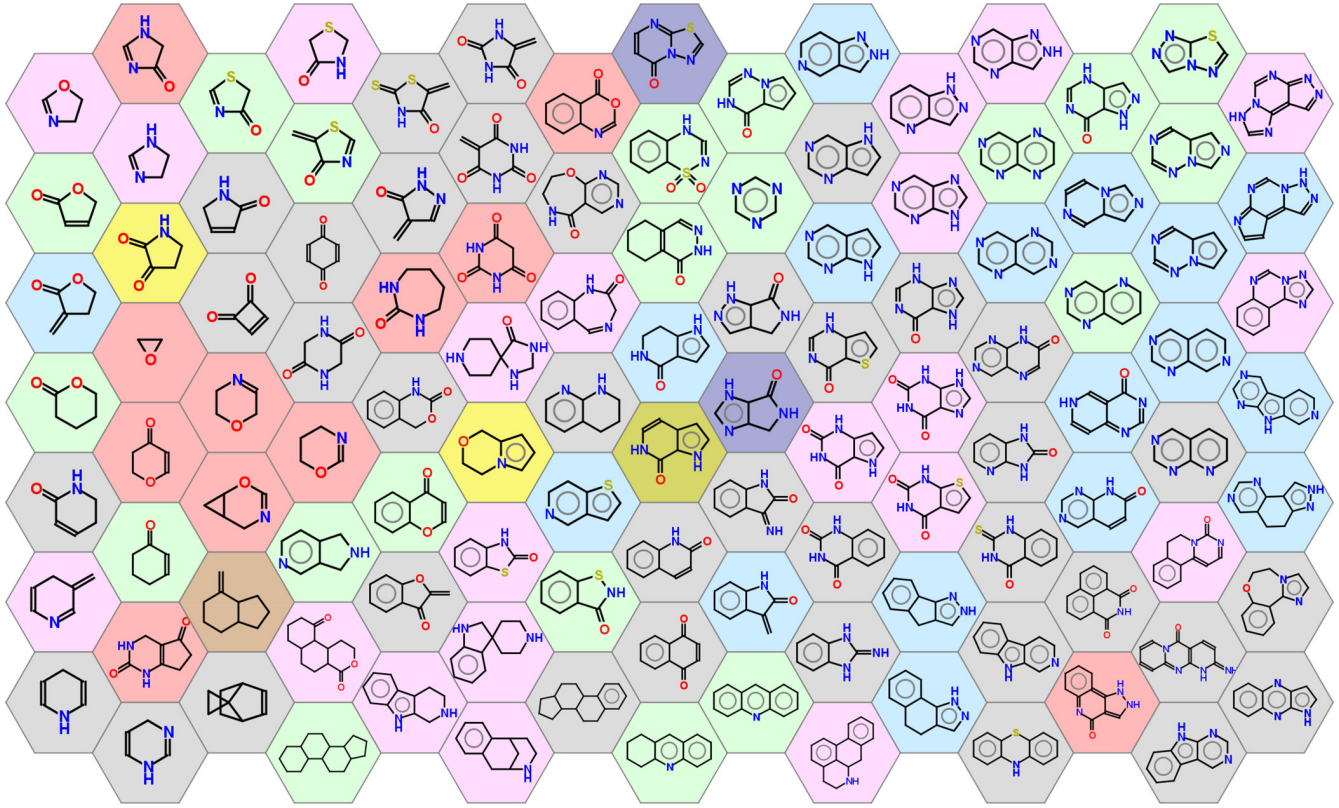
query molecule
aligned, nearest reference molecule

 Chen Y. et al., J Chem Inf Model 2020, 60, 2858–2875.

Bioactive ring systems

- The **vast majority of small-molecule drugs** (>90%) **contain at least one ring system**, regardless of whether they are of natural or synthetic origin or not[1]

- Ring systems form the **structural core** of most molecules and **determine their shape and conformational flexibility, as well as the orientation of substituents**[2]
  → they are often essential to biological activity[3]



Most common bioactive rings.[3] The preferred target class: GPCRs, kinases, proteases, other enzymes, nuclear receptors, ion channels, epigenetic targets, other targets, multiple targets.

1. R. D. Taylor, M. MacCoss and A. D. G. Lawson, J. Med. Chem., 2014, 57, 5845–5859.
2. P. Ertl, S. Jelfs, J. Mühlbacher, A. Schuffenhauer and P. Selzer, J. Med. Chem., 2006, 49, 4568–4573.
3. P. Ertl, J. Chem. Inf. Model., 2022, 62, 2164-2170.

Cheminformatics studies on natural product ring systems

- Lee and Schneider in 2001[1]
  - 10,495 NPs
  - 17% of these ring systems are represented in a compound collection of world trade drugs
- Ertl and Schuffenhauer in 2008[2]
  - 113,664 unique molecules extracted from the Dictionary of Natural Products[3]
  - NP ring systems form a highly diverse, feature-rich pool of structural templates for library and compound design

- Limitations of previous studies
  - **Known NP space is expanding quickly** → over 250,000 known NPs[4]
  - Largely disregard key molecular properties related to **stereochemistry, 3D shape and electrostatics**
- **Our aim: most accurate statistics on the ring systems from known natural products**

1. M. L. Lee and G. Schneider, J. Comb. Chem., 2001, 3, 284–289.
2. P. Ertl and A. Schuffenhauer, in Natural Compounds as Drugs, eds. F. Petersen and R. Amstutz, Birkhäuser, Basel, 1st edn., 2008, vol. 66, pp. 217–235.
3. Dictionary of Natural Products, https://dnp.chemnetbase.com
4. Y. Chen, C. de Bruyn Kops and J. Kirchmair, J. Chem. Inf. Model., 2017, 57, 2099–2111.

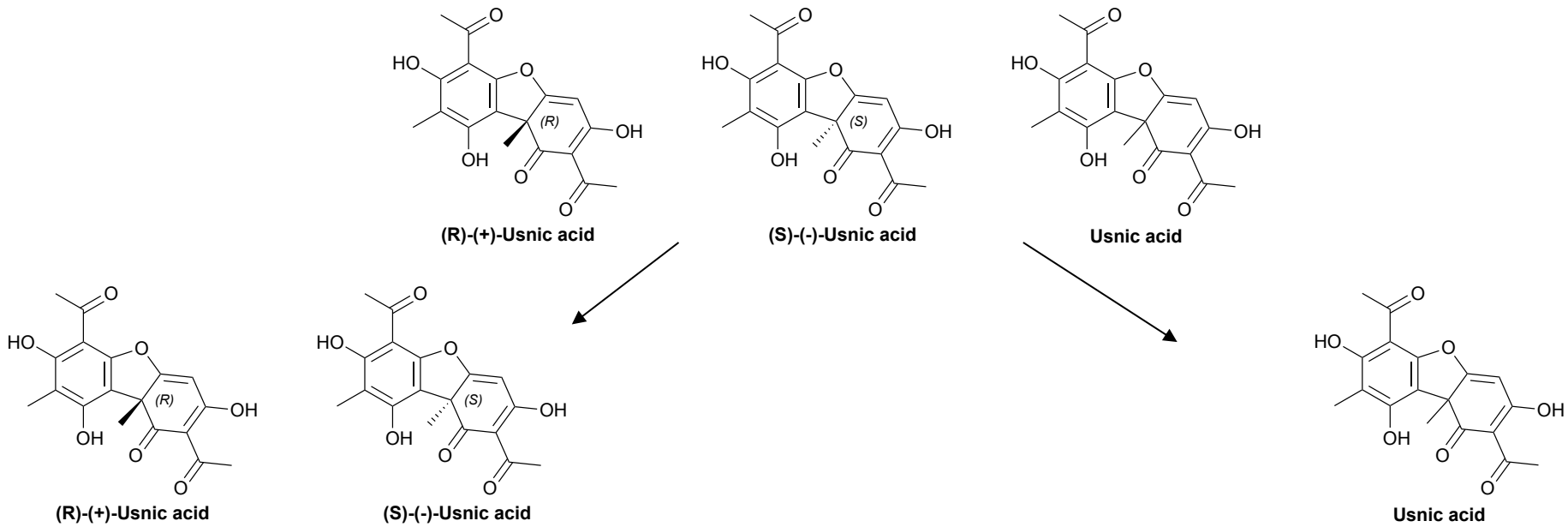
Data sets

- Natural products (NPs)
  - Curated from the Collection of Open Natural Products (COCONUT) database[1]
  - Remove data sources that contains non-neglectable portion of synthetic compounds
    - Indicated in original databases' publications and websites
    - Visual inspection of compounds included in the individual data sets especially those flagged by NP-Scout[2] as being likely of synthetic origin
    - Targeted searches for molecules with substructures characteristic to synthetic compounds: polyhalogenated alkyl chains, sulfonamides and thioureas
  - Extracted subsets of NPs from plants, bacteria, fungi and marine life
- Synthetic compounds (SCs)
  - Curated from "in stock" subset of the ZINC20 database[3]
  - Remove overlaps with "biogenic" subset
  - Remove overlap with the complete COCONUT database
- Approved drugs
  - "Approved" subset of DrugBank[4]

1. M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik and C. Steinbeck, J. Cheminform., 2021, 13, 2.
2. Y. Chen, C. Stork, S. Hirte and J. Kirchmair, Biomolecules, 2019, 9, 43.
3. J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, et al., J. Chem. Inf. Model., 2020, 60, 6065–6073.
4. D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu,et.al.et al, Nucleic Acids Res., 2018, 46, D1074–D1082.

Methods

- **Ring system extraction**: all atoms forming one or more rings (i.e., including fused and spiro rings), plus any proximate exocyclic atom connected to the ring atom via any bond other than a single bond
- Stereochemical information is often incomplete in the databases (and sometimes even wrong): New approach to maximize the use of the available stereochemical information was deployed
- https://github.com/anya-chen/RingSystems



…when considering stereochemical information: evidence-based logic

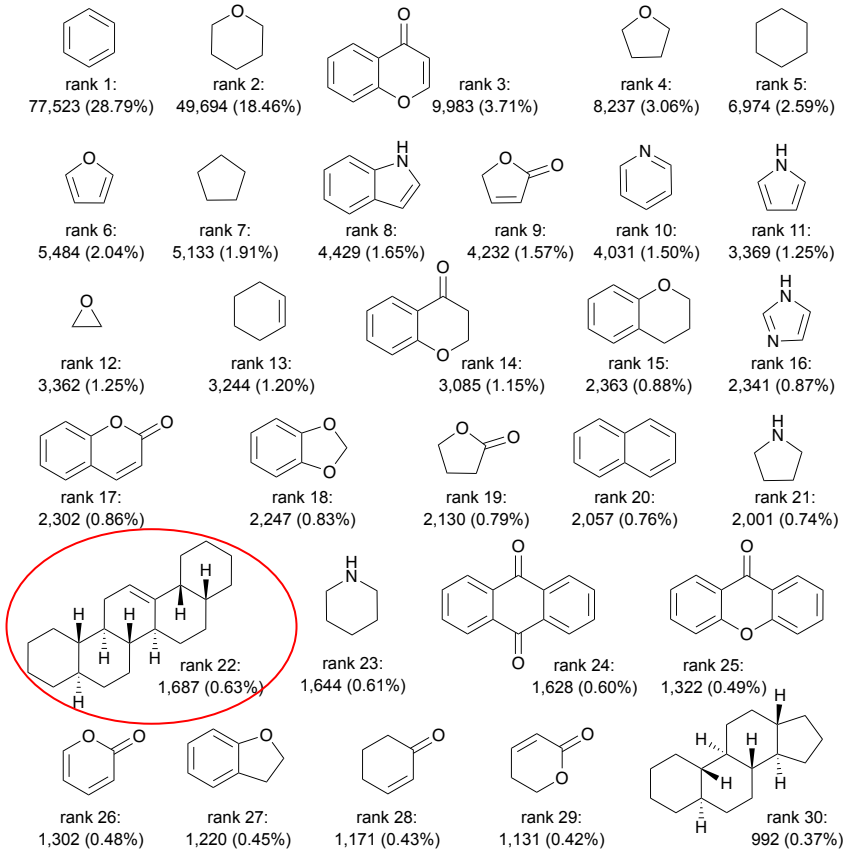
…when disregarding stereochemical information

 Chen Y et al. Nat Prod Rep 2022, DOI 10.1039/d2np00001f

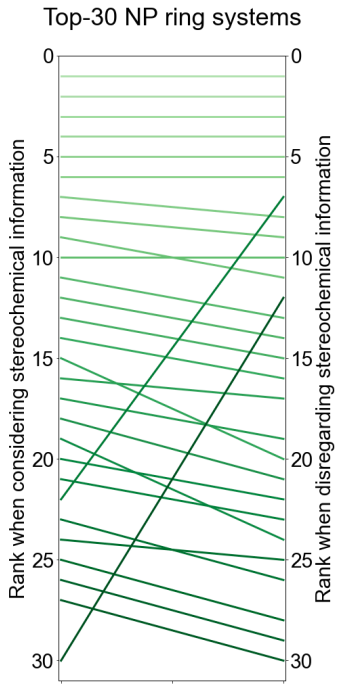
Diversity of ring systems

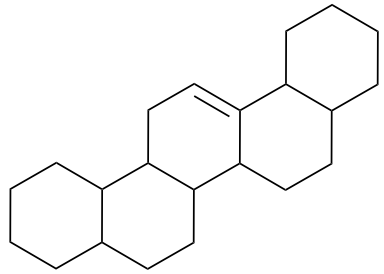
	No. unique compounds	No. unique ring systems	No. compounds/no. ring systems
When considering stereochemical information			
Natural products	**269 226**	**38 662**	6.96
Synthetic compounds	8 790 153	**53 229**	165.14
Approved drugs	2 238	602	3.72
When disregarding stereochemical information			
Natural products	246 320	**31 003**	**7.95**
Synthetic compounds	6 312 695	**30 265**	**207.41**
Approved drugs	2225	596	3.73

Most frequent ring systems



The 30 most frequent ring systems in NPs



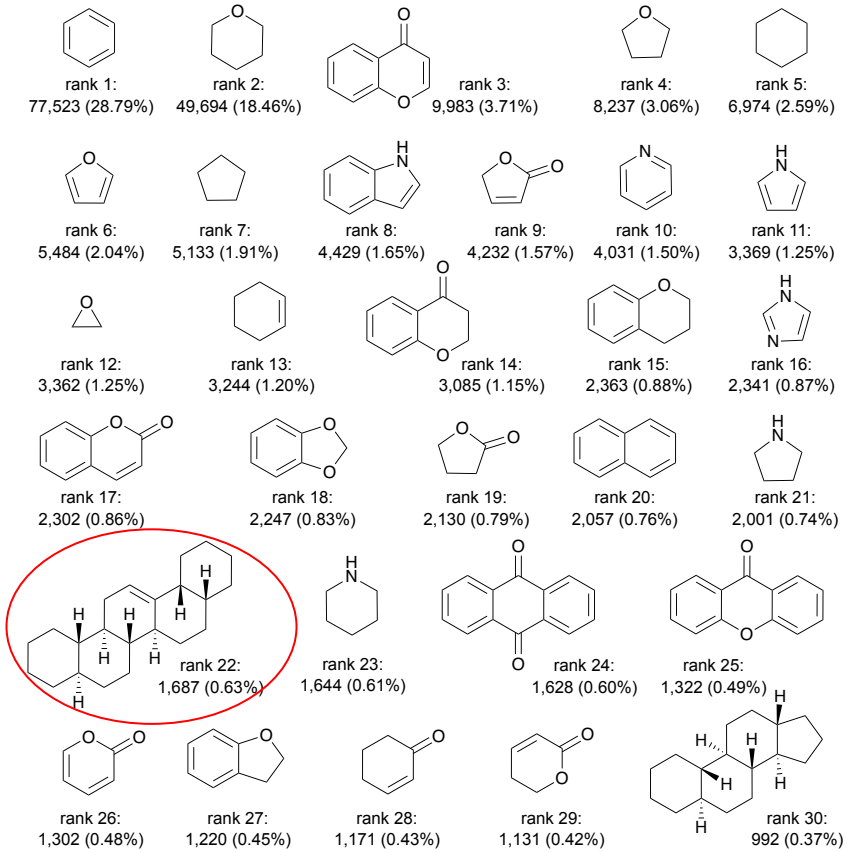


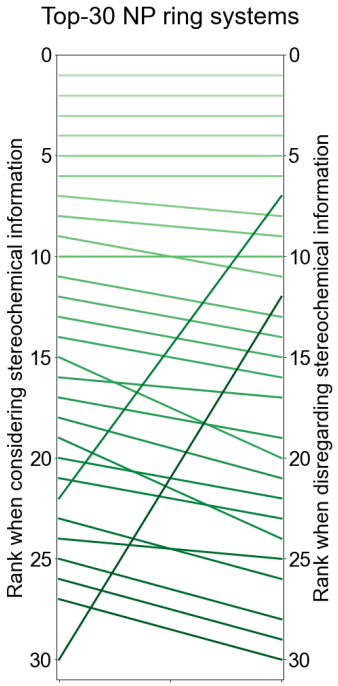
rank 7:
2.05%

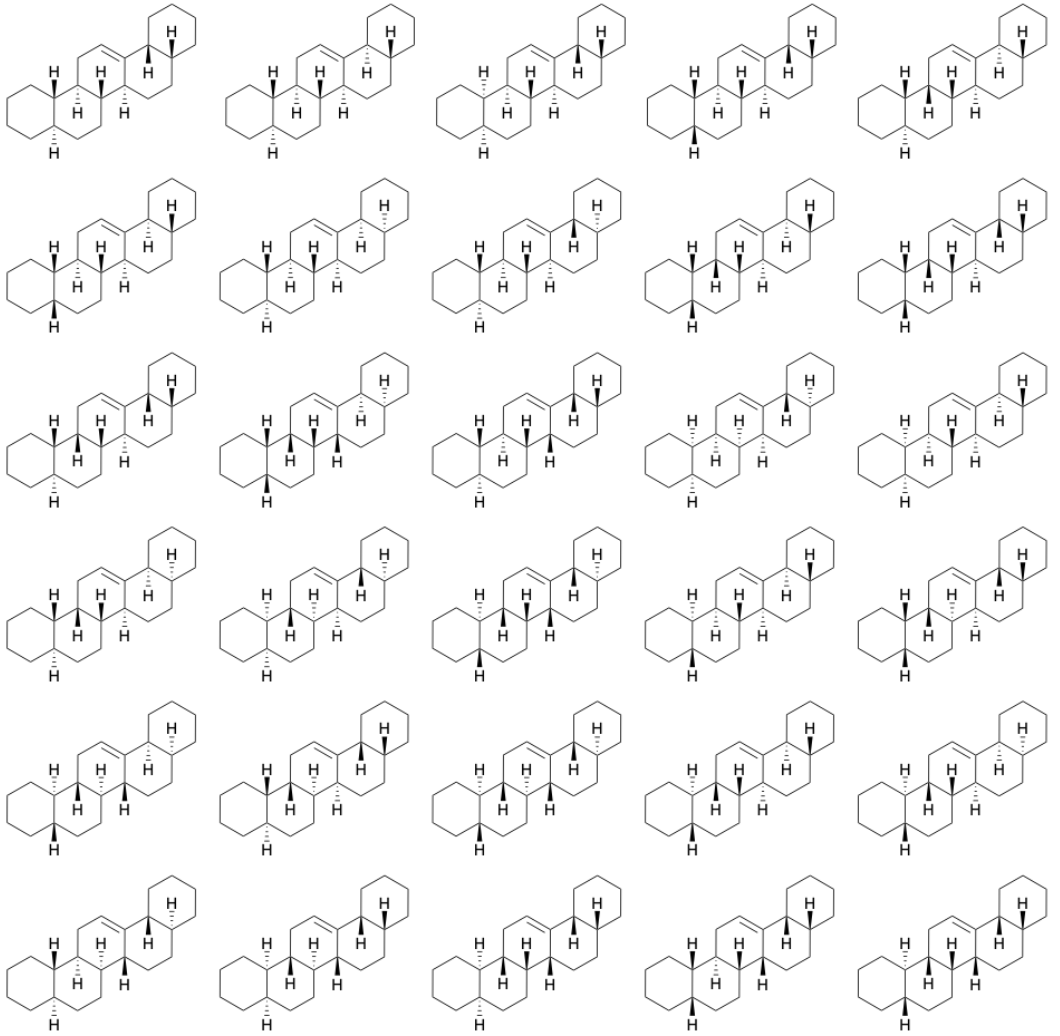
ursolic acid
7 tetrahedral atoms
105 stereoisomers recorded in databases

 Chen Y et al. Nat Prod Rep 2022, DOI 10.1039/d2np00001f

Most frequent ring systems







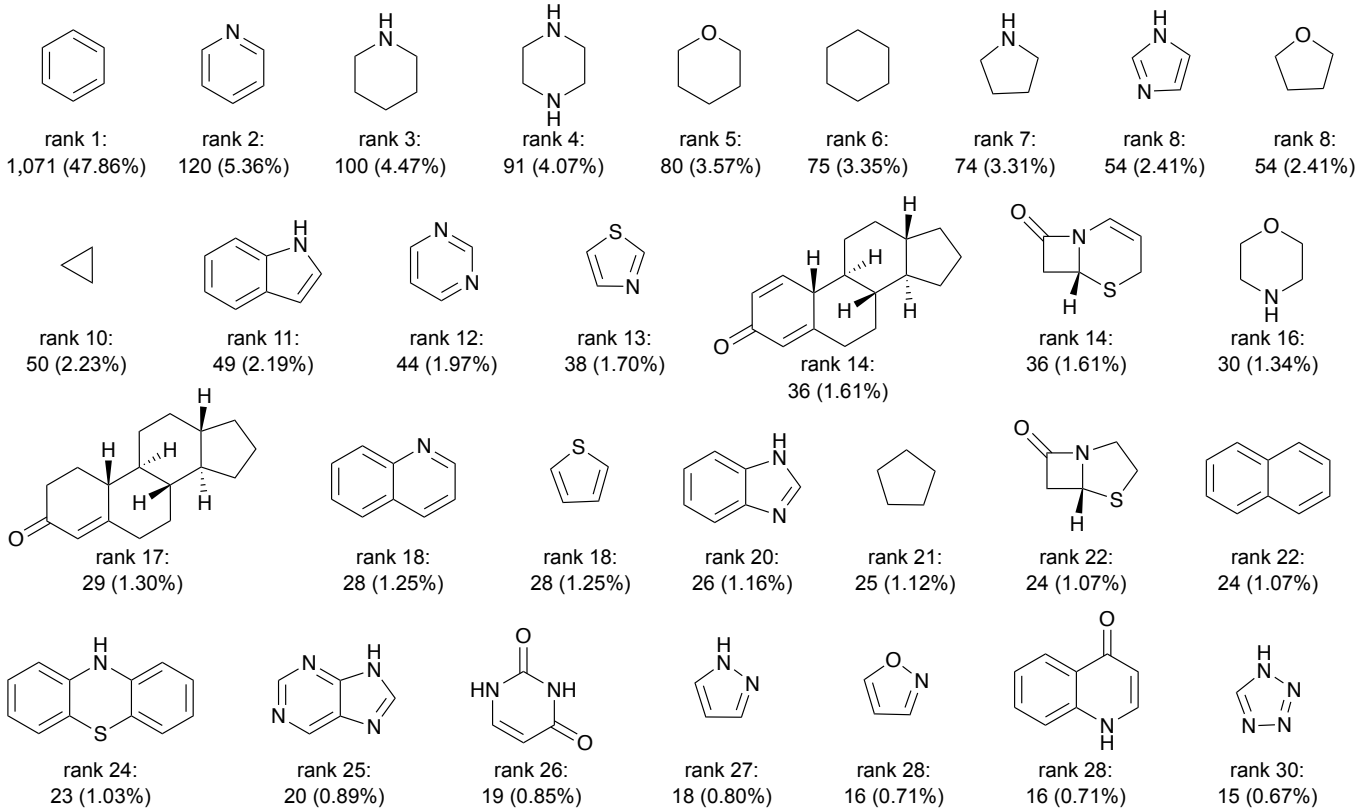
The 30 most frequent ring systems in NPs

The 30 most frequent stereoisomers

 Chen Y et al. Nat Prod Rep 2022, DOI 10.1039/d2np00001f

Natural product ring systems in approved drugs

- Of the 602 ring systems present in the approved drugs, 426 (71%) are present in natural products
- Only about 2% of the ring systems observed in natural products are represented in the approved drugs

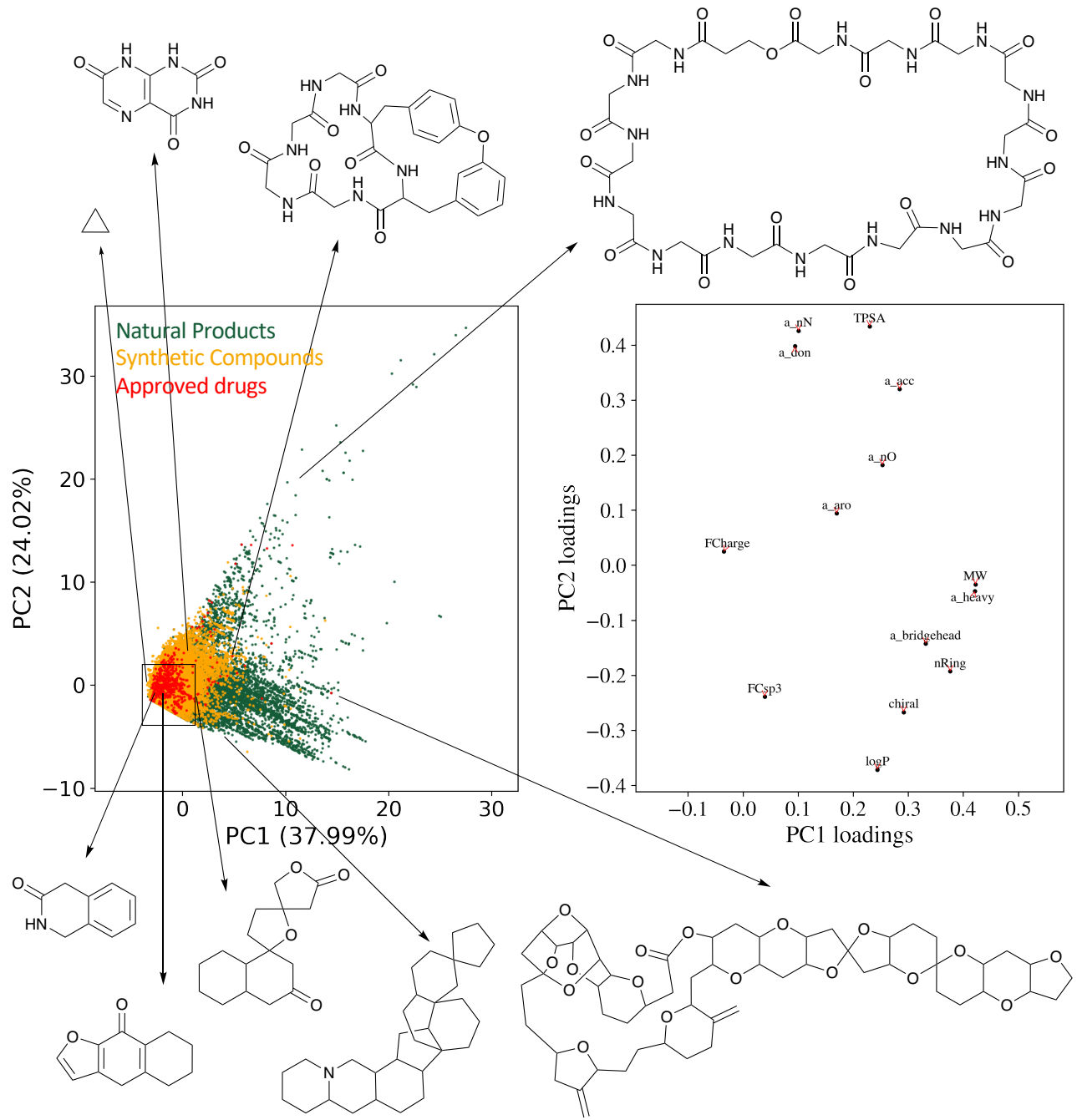


The 30 most frequent ring systems present in approved drugs

 Chen Y et al. Nat Prod Rep 2022, DOI 10.1039/d2np00001f

Physicochemical properties

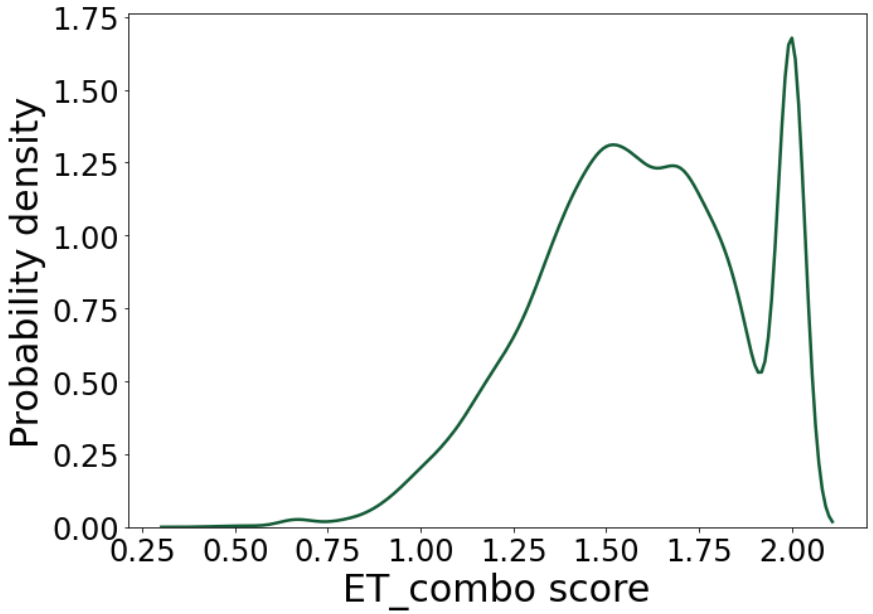
- The NP ring systems populate a wider chemical space than those derived from SCs
- The area most densely populated with NP and SC ring systems alike is also the one that is of primary relevance to small-molecule drug discovery



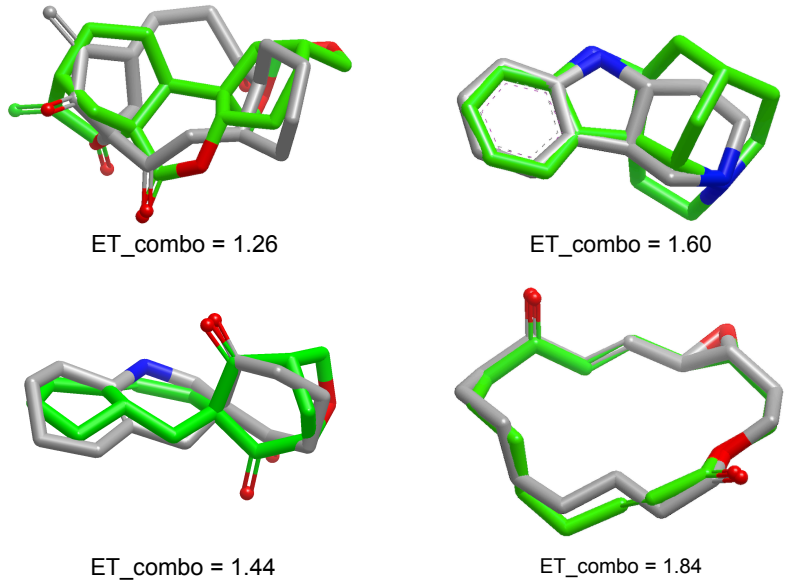
 Chen Y et al. Nat Prod Rep 2022, DOI 10.1039/d2np00001f

3D shape and electrostatic properties of ring systems

- ~15% of NP ring systems represented by identical or closely related ring systems in the SC data set
- ~50% of NP ring systems (~13,500) “covered” by a ring system of the SC data set



The maximum pairwise similarities calculated for each NP ring system and its nearest neighbor in the set of ring systems derived from the SC data set.



Alignments with different ET_combo scores

 Chen Y et al. Nat Prod Rep 2022, DOI 10.1039/d2np00001f

universität wien

Summary

- Structures of NP ring systems are **much more diverse** than those of ring systems observed in synthetic compounds
- Only about 2% of the NP ring systems are observed in approved drugs, leaving **a huge number of potential ring systems to be explored** in small-molecule drug discovery
- Approximately **half of the NP ring systems are represented by ring systems** in synthetic compound with closely related 3D shape and electrostatic properties

Natural Product Reports

ROYAL SOCIETY OF CHEMISTRY

HIGHLIGHT

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d2np00001f

Ring systems in natural products: structural diversity, physicochemical properties, and coverage by synthetic compounds†

Ya Chen, *[a] Cara Rosenkranz,[b] Steffen Hirte [ac] and Johannes Kirchmair [a]

Covering: up to 2021

The structural core of most small-molecule drugs is formed by a ring system, often derived from natural products. However, despite the importance of natural product ring systems in bioactive small molecules, there is still a lack of a comprehensive overview and understanding of natural product ring systems and how their full potential can be harnessed in drug discovery and related fields. Herein, we present a comprehensive cheminformatic analysis of the structural and physicochemical properties of 38 662 natural product ring systems, and the coverage of natural product ring systems by readily purchasable, synthetic compounds that are commonly explored in virtual screening and high-throughput screening. The analysis stands out by the use of comprehensive, curated data sets, the careful consideration of stereochemical information, and a robust analysis of the 3D molecular shape and electrostatic properties of ring systems. Among the key findings of this study are the facts that only about 2% of the ring systems observed in NPs are present in approved drugs but that approximately one in two NP ring systems are represented by ring systems with identical or related 3D shape and electrostatic properties in compounds that are typically used in (high-throughput) screening.

Received 4th January 2022

DOI: 10.1039/d2np00001f

rsc.li/npr

1. Introduction

Natural products (NPs) have a long record of use in traditional medicines. They also remain one of the most prolific sources of inspiration for modern small-molecule drug discovery.[1,2] According to the latest survey of Newman and Cragg on the origin of approved drugs,[3] 68% of all small-molecule drugs approved between 1981 and 2019 are NPs, NP derivatives, NP mimics, or structures containing NP pharmacophores.

NPs are, on average, heavier and more hydrophobic than synthetic compounds explored in the context of drug discovery.[4–6] They also feature a higher content of oxygen atoms and a lower content of nitrogen atoms.[4,5] Most outstanding, however, is their enormous structural diversity and, in part, high molecular complexity.[5–7] In particular the stereochemical properties of NPs can pose fundamental challenges to organic chemistry.

Due to the difficulties involved in the sourcing and synthesis of NPs, the availability of materials for experimental testing is limited.[8] In a recent survey of more than 250 000 known NPs we found that only approximately 10% are readily obtainable from commercial and non-commercial sources.[9] Experimental high-throughput screening (HTS) therefore rarely is an option in NPs research. Instead, a strategy which has been applied very successfully in the search for novel, bioactive NPs is virtual screening.[10] The power of virtual screening methods lies in their capacity to cherry-pick the few, most promising compounds for sourcing and testing, thereby enabling researchers to optimise the use of the limited experimental resources. Examples include the identification of influenza neuraminidase inhibitors with docking[11] and shape-based approaches,[12] the discovery of

[a]Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria. E-mail: ya.chen@univie.ac.at

[b]Center for Bioinformatics (ZBH), Universität Hamburg, 20146 Hamburg, Germany

[c]Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, 1090 Vienna, Austria

† Electronic supplementary information (ESI) available: Details on the computational methods and how to access the source code; Table S1, reporting the full names of the data sources of the COCONUT database; Table S2, reporting the numbers and percentages of NP ring systems that are matched by a ring system in the SC data set at different cutoffs of the ET_combo score. Fig. S1, showing the occurrences (in percent) of the 30 most frequent ring systems in (a) NPs and (b) SCs when considering stereochemical information; Fig. S2, showing the 30 most frequent stereoisomers of the pentacyclic triterpene ranked no. 7 of the NP ring system set when disregarding stereochemical information; Fig. S3, showing every 500th (a) NP ring system and (b) SC ring system (stereochemical information considered; singletons omitted); Fig. S4, showing the 30 most diverse (a) NP ring systems and (b) SC ring systems (identified by a *k*-means clustering method implemented using scikit-learn and RDKit that takes Morgan2 fingerprints with a length of 1024 bits as input; singletons removed prior to clustering); Fig. S5, showing the 35 ring systems recorded for at least 20 times in each of the subsets of NPs from plants, bacteria, fungi and marine life. See https://doi.org/10.1039/d2np00001f

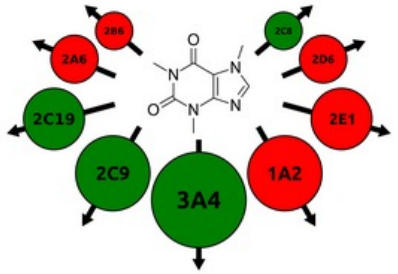
NERDD

New E-Resource for Drug Discovery

nerdd.univie.ac.at

Stork et al., *Bioinformatics* 2019, 36, 1291–1292

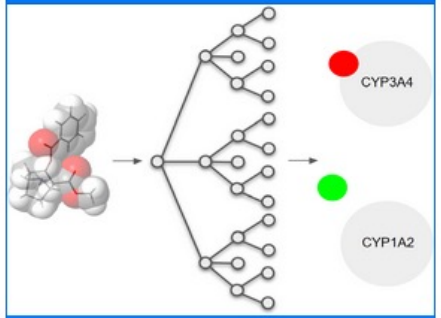
Cytochrome P450 substrates



CYPstrate

Prediction of Cytochrome P450 substrates

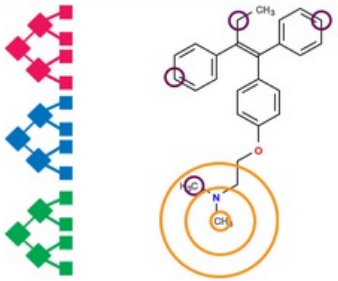
Cytochrome P450 inhibitors



CYPlebrity

Prediction of Cytochrome P450 inhibitors

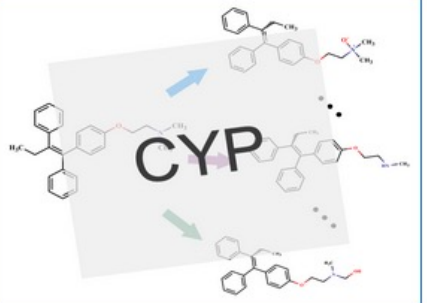
Sites of Metabolism



FAME 3

Regioselectivity prediction for phase 1 and phase 2 metabolism

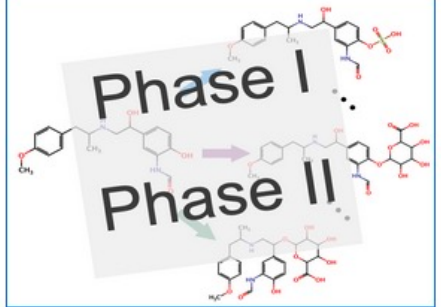
Metabolite Structures



GLORY

Metabolite structure prediction for cytochrome P450 metabolism

Metabolite Structures



GLORYx

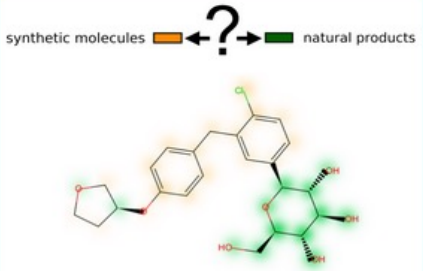
Metabolite structure prediction for phase I and II metabolism

Frequent Hitters

Hit Dexter 3

Prediction of frequent hitters

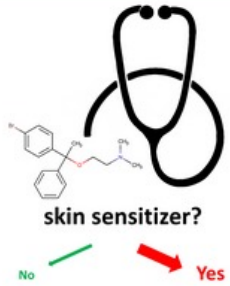
Natural Product-Likeness



NP-Scout

Identification and visualization of natural product-likeness

Skin Sensitization



Skin Doctor CP

Prediction of skin sensitization potential