Electronic Supplementary Material (ESI) for Natural Product Reports.

Supporting information

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

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

[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

Methods

Data sets of natural products

A comprehensive, clean set of known natural products (NPs) was derived from the Collection of Open Natural Products (COCONUT) database.[1] First, the COCONUT database was extracted from the SourceNaturalProduct collection of the COCONUT mongodb dump[2] (version 2020-10). Then, any entries (i.e. molecules) originating from any of the following data resources removed because they contain a non-neglectable portion of synthetic compounds (SCs): “bitterdb”, “chembl_np”, “chemspidernp”, “conmednp”, “exposome-explorer”, “fooddb”, “gnps”, “ibs2019mar_nc”, “indofinechemical”, “lichendatabase”, “specsnp”, “supernatural2”, “swmd” or “zincnp” (see Table S1 for a description of all sources). The impure data sets were identified based on (i) information provided on the websites of the individual data resources, (ii) information provided in scientific publications accompanying individual data sets (iii) visual inspection of the compounds included in the individual data sets, in particular those compounds flagged by NP-Scout[3,4] as being likely of synthetic origin, and (iv) targeted searches for molecules with substructures characteristic to SCs: polyhalogenated alkyl chains, sulfonamides and thioureas (as exemplified in ref. 5).

Based on the refined NP data set, subsets representing NPs from individual kingdoms were compiled (using the information provided in the “organism Text” field). More specifically, the subset of NPs from plants was compiled from "afrocancer", "afrodb", "afromalariadb", "biofacquim", "cmaup", "etmdb", "himdb", "hitdb", "inpacdb", "knapsack", "mitishamba", "npact", "p-anapl", "respect", "spektraris", "tcmid", "tipdb", "tppt" and "vietherb". Not included in the plants subset are NPs from the “nanpdb”, “sancdb”, “tcmdb_taiwan”, and “uefs” because not all NPs from these sources are produced by plants (even though they are annotated as plant-based NPs in the COCONUT database).

The subset of NPs from bacteria was extracted from "knapsack", "npatlas", "np_atlas_2019_12", "piellabdata", "streptomedb" and "streptomedb3". Analogously, the subset of NPs from fungi was extracted from "npatlas", "np_atlas_2019_12", "biofacquim" and "knapsack". The subset of NPs from marine species was composed of all compounds from “cmnpd” and “mnp”.

The SMILES notations of all the compounds (“originalSmiles” field) were converted into Molecule objects with RDKit.[6] The atom indices that are part of the original SMILES notations were removed with RDKit. Then, the structures were standardised with the ChEMBL Structure Pipeline.[7,8] This procedure involved the removal of salts and the

neutralisation of formal charges. Compounds with elements other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br and I were removed. Multi-component compounds for which the ChEMBL Structure Pipeline did not unambiguously identify the major component were split into individual compounds. Any molecules containing the substructures “[*;R]=[*;R]=[*;R]” (two double bonds originating from the same ring atom) and “[#6+]” (a positively charged carbon atom) were removed. Then, for each data set, deduplication of the molecules based on canonical SMILES was performed.

Data set of synthetic compounds

In order to compile a comprehensive, clean set of readily purchasable SCs, the "in-stock" subset of the ZINC20[9,10] database was downloaded. This nine-million-compounds data set is polluted by a small number of NPs. In order to identify these NPs, the "in-stock" subset was overlaid with the "biogenic" subset (which includes natural products, derivatives of natural products, and metabolites), also from the ZINC20 database, and the NPs were removed based on ZINC IDs.

The molecular structures (represented by SMILES notations) were prepared using the identical protocol described for the NPs. In order to obtain an as-clean-as-possible data set of SCs, any compounds present also in the complete COCONUT database were removed (based on the canonical, non-isomeric SMILES generated from tautomers standardised with the TautomerEnumerator class of RDKit; these reduced representations were used for this deduplication procedure only).

Data set of approved drugs

The "Approved" subset of DrugBank[11,12] was retrieved from the online repository. The molecular structures were prepared following the identical protocol used for the preparation of the NPs.

Identification of unique molecular structures

In the scenario disregarding stereochemical information, unique molecular structures were identified based on the non-isomeric SMILES notations generated with RDKit.

In the scenario considering stereochemistry (i.e. tetrahedral atom configuration), pairs of molecules were tested for identity according to a procedure that builds on the evidence-based approach described in the introductory section of Results and Discussion. The procedure returns TRUE for a pair of molecules, m1 and m2, if the two molecules are identical (more accurately, if there is no evidence that the molecules are not identical):

- If the constitution of m1 and m2 is distinct (based on their SMILES notations, with any stereochemical information removed):
    - return FALSE
- If the constitution of m1 and m2 is identical (based on their SMILES notations, with any stereochemical information removed):
    - Generate all possible substructure matches between m1 and m2 (using the *GetSubstructMatches* function of RDKit; stereochemical information disregarded with *useChirality=False*)
    - For each substructure match:
        - For each pair of matching atoms:
            - If the configuration of exactly one atom is not specified:
                - Add the unspecified atom to *unspecified_atoms* (a list of atoms for which their configuration will be enumerated)
        - Enumerate all possible enantiomers based on all atoms in *unspecified_atoms* (this results in 2^n enantiomers, where *n* is the number of atoms in *unspecified_atoms*)
        - For each enantiomer:
            - Test whether m1 and m2 can be superposed (with the *HasSubstructMatch* function in RDKit; this time with *useChirality=True*)
                - If yes:
                    - return TRUE
    - return FALSE

This algorithm can be directly accessed via https://github.com/anya-chen/RingSystems/tree/master/RingSystems/RingSystemClass.py.

Definition and identification of ring systems

For the purpose of this study, and analogous to previous studies,[13–15] ring systems are defined as all atoms forming a ring, plus any proximate exocyclic atom(s) connected via any type of bond other than a single bond. Two rings sharing at least one atom (i.e. fused and spiro rings) are considered as one ring system.

In order to obtain the ring systems the following algorithm was applied to each chemical structure with at least one ring:

1. Split of molecule into individual rings (with the RDKit function *ringInfo*). This process results in one or more ring atom sets.
2. If two ring atom sets share at least one atom the sets are fused.

3. The resulting ring systems (i.e. processed ring atom sets) are extended by all atoms directly connected to the ring via any type of bond other than a single bond.
4. All other substituents are replaced by a hydrogen atom.

From a single compound more than one ring system may be derived. Multiple occurrences of a specific ring system in one and the same molecule increase the count of ring systems by 1 (Fig. 1a).

After obtaining ring systems for each compound, unique ring system structures were identified using the procedure described in the section "Identification of unique molecular structures".

Principal component analysis

Principal component analysis (PCA)[15] was conducted with the PCA function of scikit-learn,[16] based on 14 relevant physicochemical properties (calculated with RDKit and normalised using the StandardScaler function of scikit-learn): number of oxygen atoms (a_nO), number of nitrogen atoms (a_nN), number of chiral centres (chiral), number of heavy atoms (a_heavy), number of hydrogen-bond acceptors (a_acc), number of hydrogen-bond donors (a_don), number of aromatic atoms (a_aro), number of rings (nRings), number of bridgehead atoms (a_bridgehead), molecular weight (MW), Crippen logP (o/w) (logP), topological polar surface area (TPSA), formal charge (FCharge) and fraction of sp3 hybridised carbon atoms (FCsp3).

Three-dimensional comparison of shape and electrostatic properties

For all unique NP ring systems with fully defined configuration of all tetrahedral atoms, a single, low energy conformation (because they will serve as queries for the shape-based superposition) was generated with OMEGA[16,17] (applying default settings; macrocycles were prepared with the "macrocycle mode", all others ring systems with the "classic mode").

For all SC ring systems (with and without fully defined configuration; undefined configurations were enumerated with the *flipper* feature of OMEGA), up to 200 low-energy conformations in "classic mode" or up to 400 low-energy conformations in “macrocycle mode” (default setting) were generated with OMEGA (applying default settings; macrocycles were prepared with the "macrocycle mode", all others ring systems with the "classic mode").

Next, for all ring system conformers partial charges were calculated with the AM1BCCELF10 model of OpenEye's Quacpac Toolkit.[18] The AM1BCCELF10 model uses AM1 charges[19] (derived from a semi-empirical quantum-mechanic wave function) and applies bond charge correction (BCC) afterwards.[20] Then, the Electrostatic Least-interacting Functional Groups (ELF) conformer selection[21] is applied to solve issues appearing with the AM1BCC derived charges.

Shape-based superposition of the ring systems was conducted with ROCS,[22,23] using the NP ring systems as queries and the SC ring systems as "screening database" (all settings default, except for the use of the -eon_input flag to obtain the input files required for EON[24]). The shape and electrostatic properties of the ring systems were compared with EON (all settings default, except for the -charges flag set to "existing" in order to preserve the existing partial charges, and the -fixpka_query and -fixpka_dbase flags set to "false"). More specifically, for each pair of NP and SC ring systems the maximum pairwise similarity was computed with the ET_combo score. The ET_combo score puts equal weights on the shape similarity component (ShapeTanimoto computed with EON) and the electrostatic similarity component (electrostatic Tanimoto, or ET_pb, calculated using full Poisson-Boltzmann (PB) electrostatics; EON uses the PB electrostatics toolkit Zap[18] to calculate the electrostatic potential). The ET_combo score ranges from 0 to 2, with 2 indicating a perfect match of molecular structures (ring systems).

Code availability

The source code used in this analysis is available from https://github.com/anya-chen/RingSystems.

Tables

Table S1. Full names of the data sources of the COCONUT database.

"Source" field in the COCONUT database	**Database name**	**Included in this analysis**
afrocancer	AfroCancer	yes
afrodb	AfroDB	yes
afromalariadb	AfroMalariaDB	yes
analyticon_all_np	NPs from AnalytiCon Discovery	yes
biofacquim	BIOFACQUIM	yes
bitterdb	BitterDB	no
carotenoids	Carotenoids Database	yes
chebi_np	ChEBI NPs	yes
chembl_np	ChEMBL NPs	no
chemspidernp	ChemSpider NPs	no
cmaup	CMAUP (Collective Molecular Activities of Useful Plants)	yes
cmnpd	CMNPD (Comprehensive Marine Natural Products Database)	yes
conmednp	ConMedNP	no
etmdb	ETM (Ethiopian Traditional Medicine) DB	yes
exposome-explorer	Exposome-explorer	no
fooddb	FooDB	no
gnps	GNPS (Global Natural Products Social Molecular Networking)	no
himdb	HIM (Herbal Ingredients in-vivo Metabolism database)	yes
hitdb	HIT (Herbal Ingredients Targets)	yes
ibs2019mar_nc	InterBioScreen Ltd	no
indofinechemical	Indofine Chemical Company	no
inflamnat	InflamNat	yes
inpacdb	InPACdb	yes
knapsack	KNApSaCK	yes
lichendatabase	Lichen Database	no
mitishamba	Mitishamba database	yes
mnp	Marine Natural Products	yes
nanpdb	NANPDB (Natural Products from Northern African Sources)	yes

ncidb	The NCI Development Therapeutics Program (DTP) NPs	yes
np_atlas	NPAtlas	yes
np_atlas_2019_12	NPAtlas version 2019-12	yes
npact	NPACT (Naturally occurring Plant-based Anti-cancer Compound-Activity-Target database)	yes
npass	NPASS (Natural Product Activity and Species Source)	yes
npcare	NPCARE (Natural Products for Cancer Regulation)	yes
npedia	NPEdia	yes
nubbe	NuBBEDB	yes
p-anapl	p-ANAPL (The pan-African natural products library)	yes
phenolexplorer	Phenol-explorer	yes
phytolab	PhytoLab	yes
piellabdata	Manually selected molecules	yes
pubchem_tested_np	PubChem NPs	yes
respect	ReSpect	yes
sancdb	SANCDB (the South African Natural Compounds Database)	yes
specsnp	Specs Natural Products	no
spektraris	Spektraris NMR	yes
streptomedb	StreptomeDB	yes
streptomedb3	StreptomeDB version 3	yes
supernatural2	Super Natural II	no
swmd	Seaweed Metabolite Database (SWMD)	no
tcmdb_taiwan	TCMDB@Taiwan (Traditional Chinese Medicine database)	yes
tcmid	TCMID (Traditional Chinese Medicine Integrated Database)	yes
tipdb	TIPdb (database of Taiwan indigenous plants)	yes
tppt	TPPT (Toxic Plants–PhytoToxins)	yes
uefs	UEFS (Natural Products Databse of the UEFS)	yes
unpd	UNPD (Universal Natural Products Database)	yes
vietherb	VietHerb	yes
zincnp	ZINC NP	no

Table S2. 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.

ET_combo score	Number of NP ring systems matched by a ring system in the SC data set	Percentage of NP ring systems matched by a ring system in the SC data set
2.0	4181	15.08%
1.9	4682	16.89%
1.8	7052	25.44%
1.7	10195	36.78%
1.6	13634	49.18%
1.5	17238	62.18%
1.4	20659	74.52%
1.3	23328	84.15%
1.2	25100	90.55%
1.1	26338	95.01%
1.0	27066	97.64%
0.9	27454	99.04%
0.8	27585	99.51%
0.7	27633	99.68%
0.6	27702	99.93%
0.5	27715	99.98%
0.4	27721	100.00%

Figures

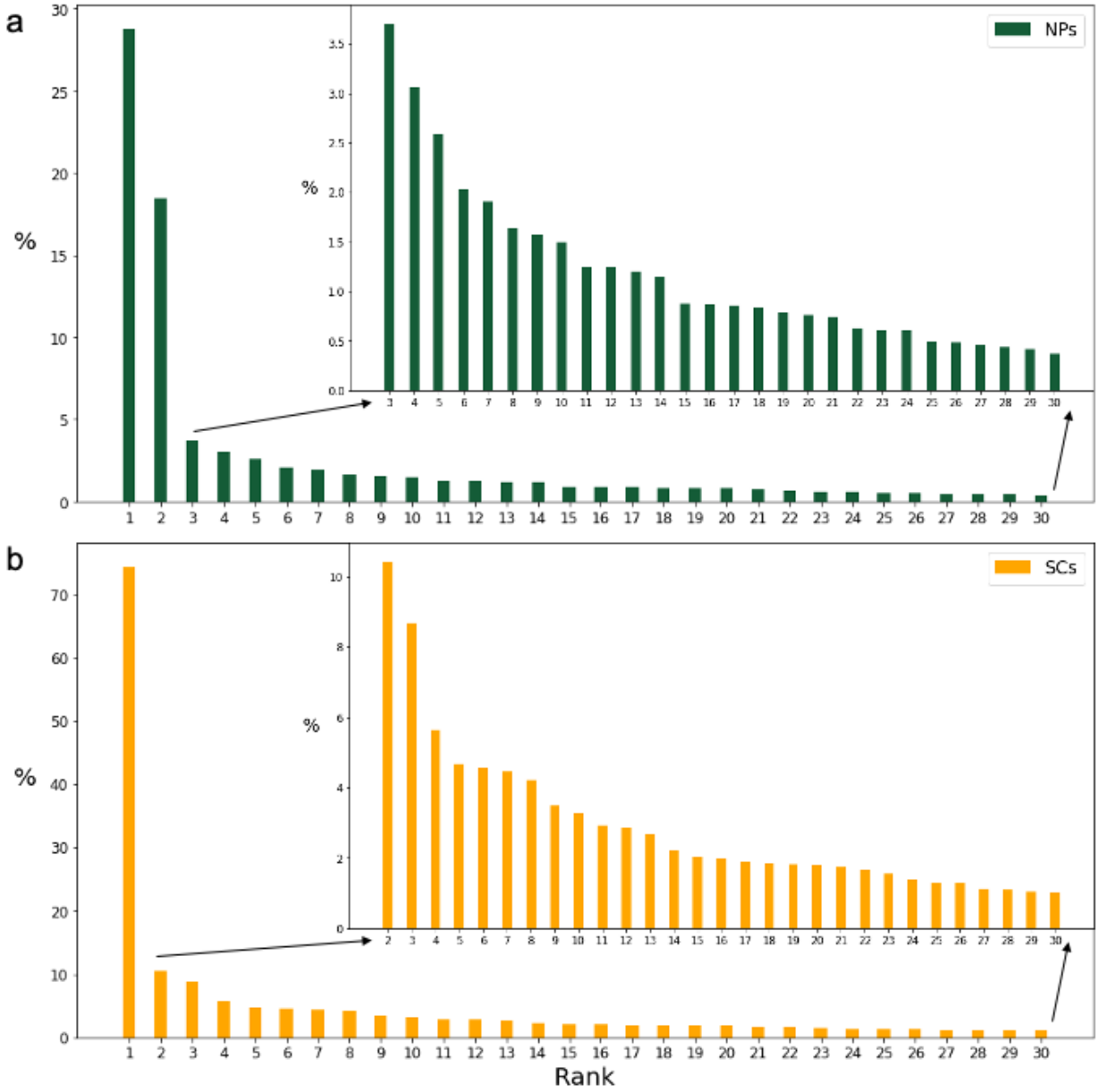


Fig. S1. Occurrences (in percent) of the 30 most frequent ring systems in (a) NPs and (b) SCs when considering stereochemical information.

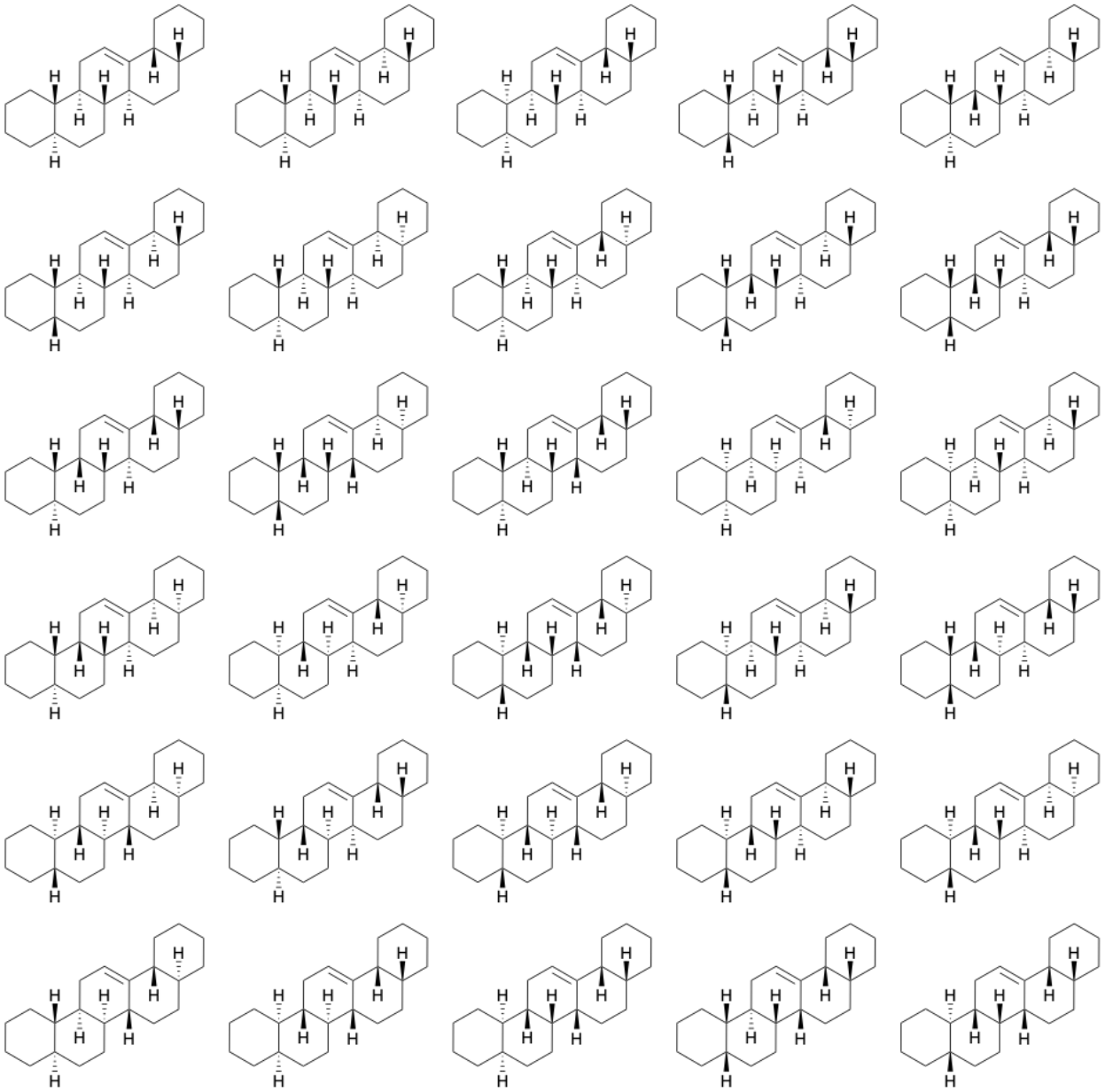
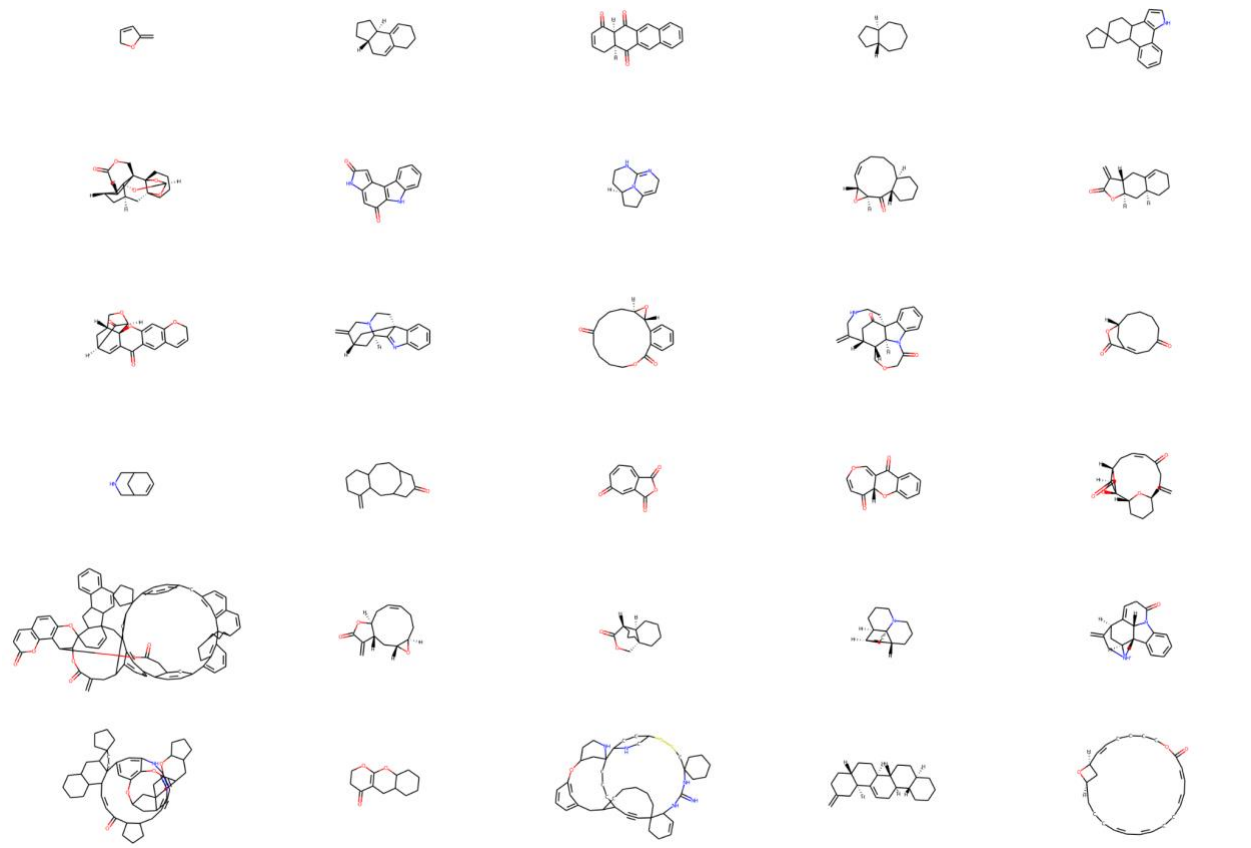


Fig. S2. The 30 most frequent stereoisomers of the pentacyclic triterpene ranked no. 7 of the NP ring system set when disregarding stereochemical information.

a



b

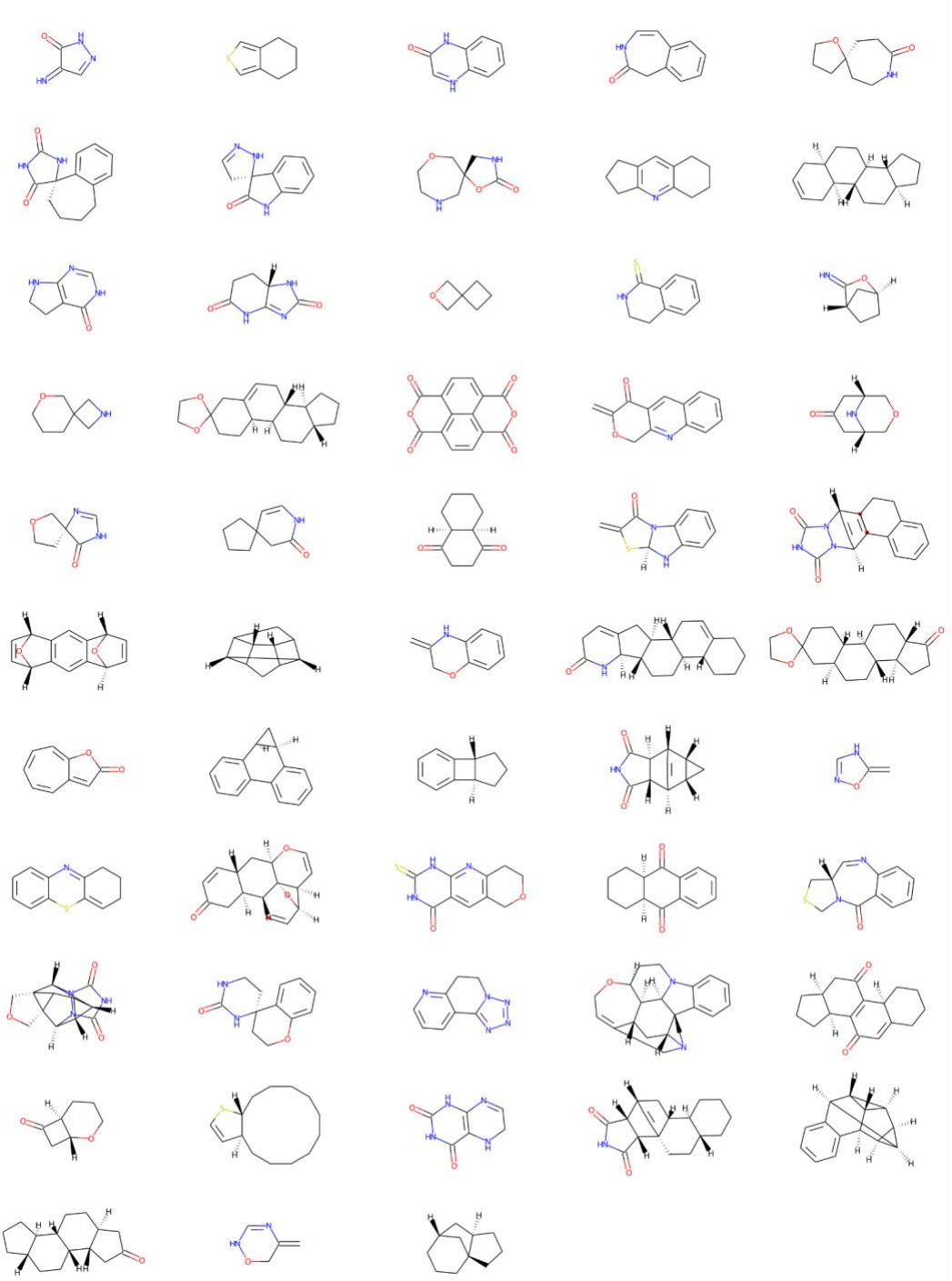


Fig. S3. Every 500th (a) NP ring system and (b) SC ring system (stereochemical information considered; singletons omitted). The structures are available also from https://github.com/anya-chen/RingSystems/tree/master/Visualization.

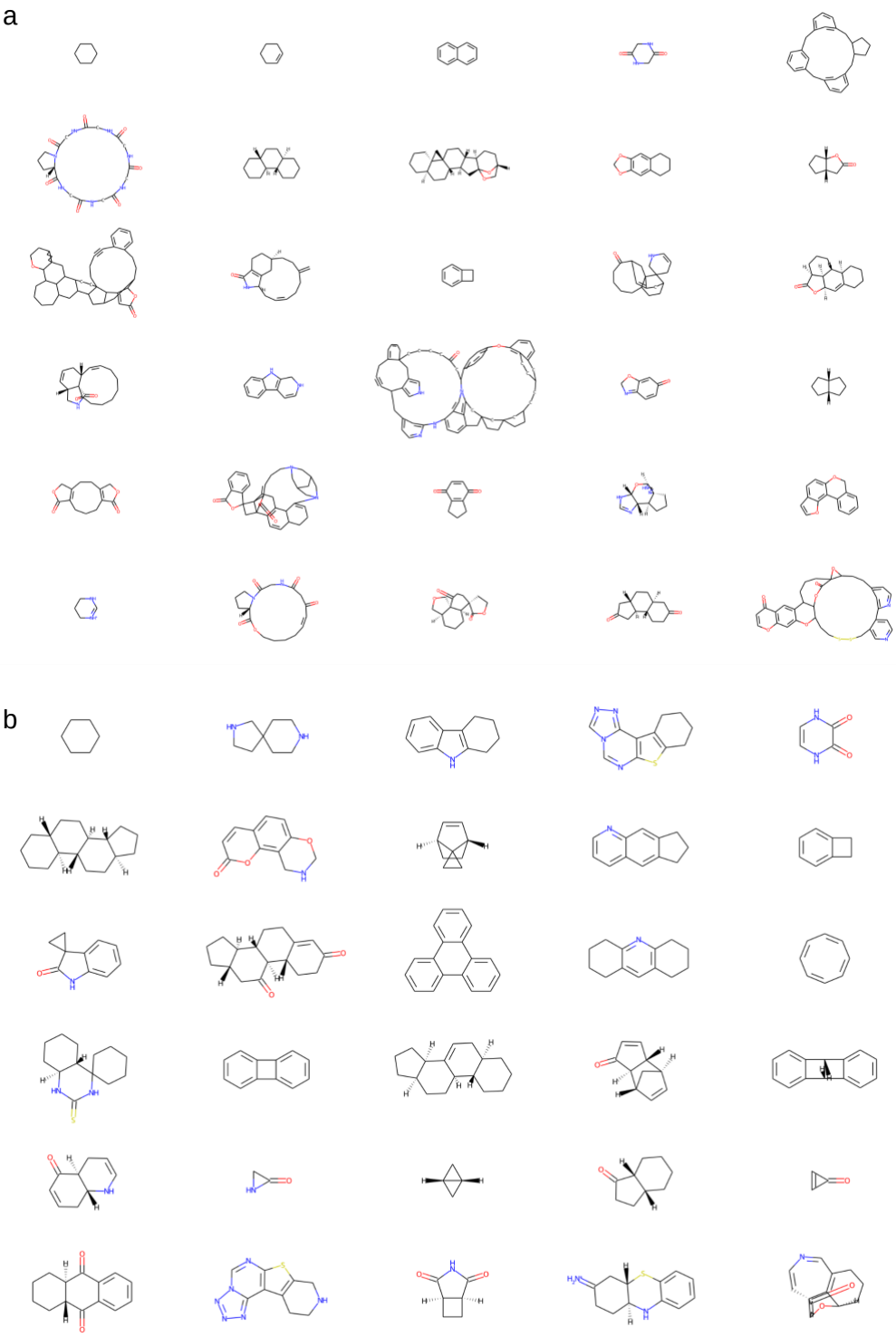


Fig. S4. The 30 most diverse (a) NP ring systems and (b) SC ring systems (identified by a k-means clustering method implemented using scikit-learn and RDKit[25] that takes Morgan2 fingerprints with a length of 1024 bits as input; singletons removed prior to clustering). The structures are available also from https://github.com/anya-chen/RingSystems/tree/master/Visualization.

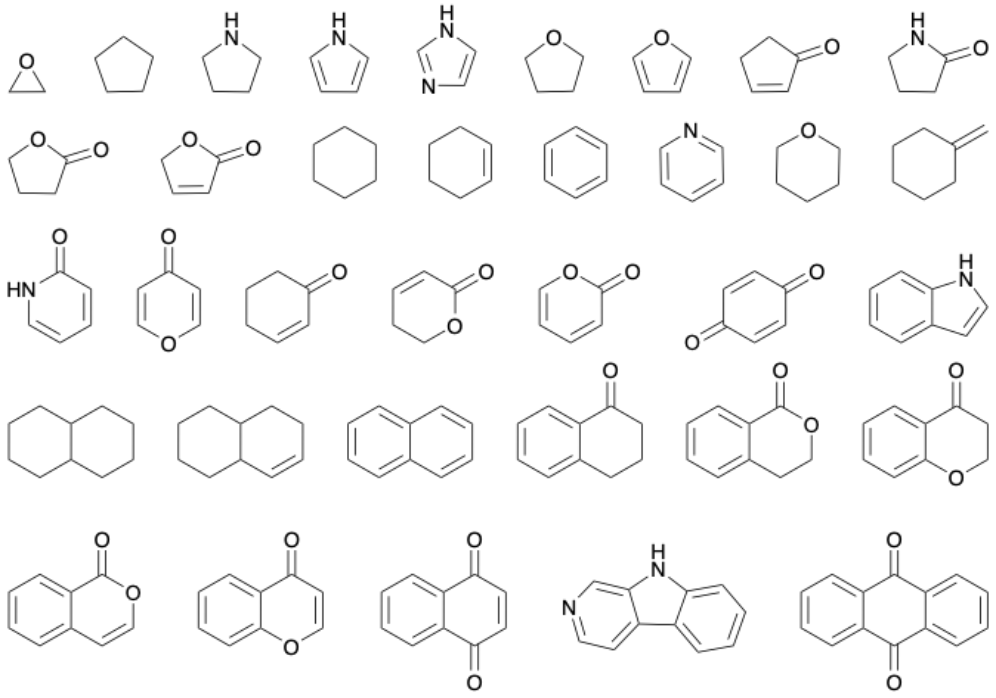


Fig. S5. The 35 ring systems recorded for at least 20 times in each of the subsets of NPs from plants, bacteria, fungi and marine life.

References

1 M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik and C. Steinbeck, *J. Cheminform.*, 2021, **13**, 2.
2 COCONUT: Natural Products Online, https://coconut.naturalproducts.net/download, (accessed 8 June 2021).
3 Y. Chen, C. Stork, S. Hirte and J. Kirchmair, *Biomolecules*, 2019, **9**, 43.
4 C. Stork, G. Embruch, M. Šícho, C. de Bruyn Kops, Y. Chen, D. Svozil and J. Kirchmair, *Bioinformatics*, 2020, **36**, 1291–1292.
5 Y. Zabolotna, P. Ertl, D. Horvath, F. Bonachera, G. Marcou and A. Varnek, *Mol. Inform.*, 2021, **40**, e2100068.
6 G. Landrum, RDKit: Open-source cheminformatics, https://www.rdkit.org, (accessed 17 September 2021).
7 A. P. Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L. J. Bellis, M. De Veij and A. R. Leach, *J. Cheminform.*, 2020, **12**, 51.
8 GitHub - chembl/ChEMBL_Structure_Pipeline: ChEMBL database structure pipelines, https://github.com/chembl/ChEMBL_Structure_Pipeline, (accessed 15 September 2021).
9 J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *J. Chem. Inf. Model.*, 2020, **60**, 6065–6073.
10 ZINC20 database, http://zinc20.docking.org/, (accessed 20 April 2021).
11 D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox and M. Wilson, *Nucleic Acids Res.*, 2018, **46**, D1074–D1082.
12 DrugBank Release Version 5.1.8, https://go.drugbank.com/releases/latest#structures, (accessed 3 December 2021).
13 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.
14 J. Hert, J. J. Irwin, C. Laggner, M. J. Keiser and B. K. Shoichet, *Nat. Chem. Biol.*, 2009, **5**, 479–483.
15 M. Aldeghi, S. Malhotra, D. L. Selwood and A. W. E. Chan, *Chemical Biology & Drug Design*, 2014, 83, 450–461.
16 P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson and M. T. Stahl, *J. Chem. Inf. Model.*, 2010, **50**, 572–584.
17 OMEGA 4.1.0.0: OpenEye Scientific Software, Santa Fe, NM, https://www.eyesopen.com, (accessed 17 September 2021).
18 OpenEye Toolkits 2020.2.2 OpenEye Scientific Software, Santa Fe, NM, https://www.eyesopen.com, (accessed 17 September 2021).
19 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902–3909.
20 A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2000, **21**, 132–146.

21 MolCharge Theory — Applications, vDev build, https://docs.eyesopen.com/applications/quacpac/theory/molcharge_theory.html, (accessed 27 December 2021).
22 P. C. D. Hawkins, A. G. Skillman and A. Nicholls, *J. Med. Chem.*, 2007, **50**, 74–82.
23 ROCS 3.4.1.0: OpenEye Scientific Software, Santa Fe, NM, https://www.eyesopen.com, (accessed 17 September 2021).
24 EON 2.3.4.0: OpenEye Scientific Software, Santa Fe, NM, https://www.eyesopen.com, (accessed 17 September 2021).
25 P. Walters, k-means clustering, https://github.com/PatWalters/kmeans, (accessed 28 April 2022)